

Sustainable outcomes from Gleneagles

African nations will be more likely to support development projects whose outcomes are indispensable to them. Participants at next week's G8 summit should focus aid in this direction.

At a summit meeting in Scotland next week, leaders of the Group of Eight (G8) industrialized nations are set to sign a deal on debt and aid that could see billions of dollars in fresh money flow into African development projects. Some of these projects will have a sizeable research component, as science and technology have been climbing up the development agenda (see the Special Report on page 1146).

The G8 leaders won't meet directly with African leaders, however: South Africa's president, Thabo Mbeki, attending as a guest, will be the continent's sole representative. This disconnect unintentionally reflects the historic divide between aid donors and recipients: the former frequently fail to consult adequately with the latter. Huge sums have been wasted because the solutions previously offered to African communities were not integrated into their everyday needs. As a result, the continent is strewn with the debris of unsuitable donated plant and equipment.

Local support

The stated priority of donors for many years now has been to focus on projects that will be sustainable when foreign support runs out. Some progress has been made in the design of such projects in sectors such as education and health. One challenge for the scientific community is to apply the same logic to development-oriented science and technology projects. Planners need to make sure that these projects are indispensable, so that local people and politicians will support them when the aid money has gone.

The approach has already been tested in some instances. In rural Kenya, for example, poor farmers in remote areas can now call on the services of paravets — local people given basic training in veterinary practice. Foreign non-governmental organizations taught and equipped interested rural people a decade ago, and the paravets now earn a living by charging farmers for their services.

"Planners need to make sure projects are indispensable, so African politicians will support them when the aid money has gone."

Donor organizations say that this service has been welcomed locally, as professional vets are in such short supply. The inclusion of traditional remedies has made farmers more willing to accept the paravets, project planners say. After the scheme became established, donors were able to take a back seat, leaving the paravets and the national veterinary service to carry the scheme forward. The project is well placed to survive without additional aid.

This kind of approach is easier in some science-related areas than others. It is hard to apply it, for example, in biomedical research. Rich nations currently support malaria research efforts in African labs, and would like to deepen the involvement of African government agencies in them. But the development of drugs or vaccines is a

lengthy process, and the other problems facing African governments — such as security, drought, malnutrition, poverty — are immediate. So malaria research is seldom seen as indispensable. Better malaria treatments are important to the future of Africa, but funding for their development is most likely to come from abroad.

Activity in other branches of science, such as meteorology, can be tied more convincingly to immediate needs. The impoverished desert nation of Chad, for example, has scant physical infrastructure and no research budget. But Samuel Mbainayel, a meteorologist in the city of N'Djamena, broadcasts his forecasts to rural farmers by radio every week. African specialists believe that his efforts could be broadened to provide valuable long-term weather forecasts, helping the government to plan — and the farmers to plant.

To get these forecasts up and running, money is needed to improve the capability of African meteorology services (see *Nature* 435, 863; 2005). Researchers have to be trained and university labs upgraded. These are things that donors are relatively good at. Once the improved forecasts are in place, and they start saving lives, aiding farmers and boosting yields, African governments are more likely to fund the system themselves. In the words of one British aid official, the meteorology services "will pay their own way".

Success on the ground

This philosophy has already borne fruit in agricultural research. Africa has had some home-grown successes in this field, such as the New Rice for Africa. Developed at the Africa Rice Center in Benin with the help of donor money, this new strain of high-yielding, drought-resistant rice is now being planted in several west African countries. Such successes have helped persuade African leaders to found an agricultural research centre in Kenya under the auspices of a continent-wide, self-supporting strategy called the New Partnership for Africa's Development.

Britain's prime minister, Tony Blair, as host of the summit, has selected Africa and climate change as the two main strands for discussion. On the latter, he can expect little progress. The United States has rebuffed all of his efforts to restart international negotiations over greenhouse-gas emissions targets, and the summit communiqué, if leaks of the draft are correct, may fail even to acknowledge the scientific reality of anthropogenic climate change, as described by the world's scientific academies on 7 June.

The summit's slim chances of success, therefore, ride on its agreeing a convincing package of additional aid and debt write-offs for Africa. Of course, if the participants really wanted to help Africa, they'd be talking about implementing their unfulfilled promises to reduce the massive agricultural subsidies that keep African farmers' produce out of rich country's markets. But debt relief and additional aid, carefully spent on sustainable projects, could still make some long-term difference to Africa's prospects. ■



Bringing neuroscience to the classroom

Is the US National Science Foundation jumping the gun with its plans for education?

Basic neuroscience and educational theory have, until now, ploughed largely separate academic furrows. But that hasn't stopped overenthusiastic individuals from designing 'brain-based' learning aids — often making healthy profits in the process.

Many of these tools have been built on gross misrepresentations of the science. Take the industry spawned by the idea that there is a 'critical period' for learning in early childhood, when the brain has the highest density of synapses. This ignores evidence that 'pruning' of synaptic connections is a necessary part of brain development.

Now the US National Science Foundation (NSF) is getting serious about the science of learning and its application in the classroom. Cognitive neuroscientists, psychologists, computer scientists and educationalists are being melded into huge collaborative teams (see page 1156).

Hopes are deservedly high. But questions need to be asked about whether the time is ripe for some of the links between basic science and educational practice that are now being proposed.

The computer-based 'cognitive tutors' being developed at Carnegie Mellon University in Pittsburgh are among the most solidly grounded aspects of the NSF initiative. The first tutor, for algebra, has already proved a boon for overstretched teachers. What's more, the Pittsburgh team's strategy for involving teachers in its ongoing research is both innovative and practical.

However, things get a little less convincing when it comes to basic

neuroscience. Researchers are planning to use magnetic resonance imaging to 'look under the hood' at the development of skills such as numeracy and reading. It's fascinating stuff, but how the results will inform educational practice remains, for now, largely a matter of speculation. Making meaningful connections between brain activity and behaviour is difficult, even under controlled lab settings.

Brain imaging is seductive, and has an unfortunate tendency to spawn breathless, overreaching media coverage. Care will be needed to ensure that these projects don't encourage ill-informed 'experts' to design yet more pseudoscientific educational tools.

That's not to say that scientific advances can't already help to inform educational policy. For instance, there is now a solid body of evidence that sleep patterns change significantly with age — and that, as a result, it makes little sense to wake teenagers up early to go to school, when their attention will be low as a natural consequence of their daily rhythms. Education authorities and schools are starting to hear this message, and some are adjusting their schedules accordingly.

There's also a strong case for putting the educational tools derived from research in neuroscience to more rigorous empirical tests. For instance, researchers who have evidence that dyslexics have problems with auditory processing have developed a program called Fast ForWord to help them learn to read. But the scientists' company is now marketing the software as a learning aid for children with no specific reading deficits, before they have gathered evidence that it helps anyone other than dyslexics. For now, providing this sort of evidence is where the emphasis should remain. ■

"How 'looking under the hood' at the development of numeracy and reading will inform educational practice remains, for now, largely a matter of speculation."

Crystal clear

Clarifying the Nature journals' policy on data deposition for chemical structures.

Everyone agrees that data that form the basis of a scientific paper need to be available to readers at the time of publication. But just how raw should such data be? And how much should be released ahead of publication to peer reviewers? These questions can get troublesome when releasing certain data can allow competitors an easy route to results for which the originating researchers have sweated blood. And nowhere is this dilemma more acute than in the data underlying the structures of biological macromolecules and of the complexes that they form.

To labs that solve the structures of proteins and other biological molecules, 'structure factor' files are like the reagents of other fields — material that can be used to enable a variety of experiments. Structure factors are the raw data from which atomic coordinates are derived. As such, they are a key aid for reviewers and readers in verifying a structure.

For many years, *Nature* and its sibling research journals have

required that crystallographers deposit the atomic coordinates of their structures in public data banks at the time of publication. In 2000 the International Union of Crystallography weighed the issues and decided that both coordinates and structure-factor files should be released upon publication, and this is now standard practice.

There is less of a consensus over how to balance the interests of authors and the needs of referees in the peer-review process: should we also require structure-factor files upon submission? After gathering the opinions and insights of a slice of the structure community, *Nature* and its sibling journals have adopted the following policy.

To help authors maintain control of their data before publication, structure-factor files are not required upon submission of a manuscript. However, editors may request them to aid in the review process. Referees who find that evaluation is dependent on factor files should contact the editor, who will obtain the necessary data set.

We believe that this policy balances appropriate control of data access before publication and the need for rigorous review. ■

"Structure-factor files are not required upon the submission of a manuscript, but editors may request them to aid the review process."

RESEARCH HIGHLIGHTS

First breath

PLoS Genetics doi:10.1371/journal.pgen.0010010 (2005)

A class of birth defects that affect the diaphragm has been linked to a faulty gene by Kate Ackerman of Harvard Medical School and her colleagues. The gene seems essential to lung development.

One in 3,000 babies are born with diaphragm defects, which often prove fatal because of associated lung problems. In mice, Ackerman's team found that mutations in the *Fog2* gene caused diaphragmatic defects and disrupted the development of lungs grown *in vitro* (pictured). They also identified a *Fog2* mutation in a human baby with defects in both organs.



PLOS/K. ACKERMAN ET AL

MOLECULAR BIOLOGY

DNA unpacked

Cell 121, 873–885 (2005)

DNA is normally packed into a dense structure called chromatin, which is stabilized by histone proteins. When a particular gene needs to be activated, nearby histones are chemically modified; this allows the DNA to be transcribed. Such changes can include the addition of acetyl and methyl groups by enzymes.

A team led by Robert Roeder of The Rockefeller University, New York, provides evidence that these modifications do not occur independently, but in coordinated combinations. They isolated a stable complex of molecules, which included both methyltransferase and acetyltransferase enzymes. The complex acts as a molecular machine to control gene transcription.

BIOTECHNOLOGY

Stocky in stature

Science doi: 10.1126/science.1113373 (2005)

The key to high-yield rice is an enzyme involved in the plant's reproduction, find researchers led by Makoto Matsuoka of Nagoya University in Japan.

Although regions of the rice genome associated with enhanced seed production were identified after the genome was first sequenced, how such sequences work was not understood. By studying transgenic plants, Matsuoka's team has discovered that one such region suppresses levels of the enzyme cytokinin oxidase, which regulates levels of a hormone involved in seed production. The team cross-bred plants with this genetic trait

with a stocky rice variety, to produce a plant (pictured below) that is heavy with seeds, but not so tall that it falls over in storms.

MICROBIAL GENETICS

Saints and sinners

Nature Biotechnol. doi:10.1038/nbt1110 (2005)

Many bacteria of the genus *Pseudomonas* are pathogens, including *P. aeruginosa*, which infects cystic fibrosis patients, and the fruit canker agent *P. syringae*. By contrast, the soil-dwelling *P. fluorescens* protects plants from pathogens and frost.

To find out how the saint differs from the sinners, researchers led by Joyce Loper from the Agricultural Research Service in Oregon and Ian Paulsen of The Institute for Genomic Research in Rockville, Maryland, sequenced the complete *P. fluorescens* genome. At 7.1 million bases, it is larger than the genomes of its less benign cousins and stuffed with gene clusters for chemicals harmful to pathogens, a third of which were previously unknown.



SOLID-STATE PHYSICS

Spin under strain

Phys. Rev. Lett. 94, 236601 (2005)

Future electronic devices could rely on the control of electronic spin, rather than charge, which means the discovery that mechanical forces can change a spin's orientation may have practical use. Scott Crooker and Darryl Smith of the Los Alamos National Laboratory in New Mexico used a standard optical technique called Kerr microscopy to picture spin flow in a semiconductor. The researchers measured the spins' response to electric, magnetic and mechanical forces, and found that a mechanical strain changed the spin independently of the electric field applied.

NEUROBIOLOGY

Clearing the mind

Nature Genet. doi:10.1038/ng1591 (2005)

Brain cells use motor proteins called dyneins to ferry harmful proteins to their doom, report researchers led by David Rubinstein of the Cambridge Institute for Medical Research, UK. This explains why protein clumps characteristic of certain kinds of motor-neuron disease form when mutations impair the dynein machinery.

The group studied the effect of such mutations in a mouse model of Huntington's disease. Mice with impaired dynein function developed the disease more quickly than mice with the Huntington mutation alone. The researchers suggest that this is because dynein propels packages of the harmful protein — in this case, mutant huntingtin — along the intracellular skeleton towards lysosomes, where they are destroyed.

IMAGE COURTESY OF M. ASHIKARI AND M. MATSUOKA

ANIMAL BEHAVIOUR

The buzzword*Curr. Biol.* **15**, 447–448 (2005)

Bumblebees take cues from each other when faced with unfamiliar flowers, report Ellouise Leadbeater and Lars Chittka from Queen Mary, University of London.

They released bees (*Bombus terrestris*; pictured) into a box containing two unfamiliar types of flower, and noted which type each bee selected. In subsequent trials, the bees were most likely to forage on flowers of the type they first chose, and usually only tried a flower of the unfamiliar variety if it was already being fed on by another bee. Because bees forage in unpredictable habitats, taking a lead from a fellow bee might maximize their nectar haul, the researchers suggest.

GENOMICS

Islands of stability*Genome Res.* doi:10.1101/gr.3577405 (2005)

Comparing the sequences of different mammalian genomes can illuminate the regions that have functional importance, because such elements tend to be conserved from one species to the next.

Arend Sidow of Stanford University in California and his colleagues demonstrate a powerful approach to performing such a comparison. They lined up sequences taken from the genomes of 29 different mammals and counted the number of substitutions of single bases that appeared along their length. Segments which carried fewer substitutions than expected by chance were considered to be 'constrained' and deemed likely to have some vital function. The team's technique, dubbed 'genomic evolutionary rate profiling',

identified constrained regions that varied from 3 to 1,000 base pairs in length.

CHEMISTRY

Cleaning up*Anal. Chem.* doi:10.1021/ac050460g (2005)

Chemists are being offered a new and improved version of the detergent they use to wash proteins from cell membranes for study. The new detergent molecule collapses to prevent it interfering with mass spectrometry measurements.

Researchers led by Richard Caprioli of Vanderbilt University in Nashville designed the detergent to fall apart after an acid treatment. A fragment is left behind that binds the membrane protein into a solid matrix ready for analysis. When the technique was applied to the entire contents of a cell, it revealed many more proteins than were recovered with a commercial detergent.

IMAGE
UNAVAILABLE
FOR COPYRIGHT
REASONS

FLUID DYNAMICS

Slip and slide*Phys. Rev. Lett.* **94**, 244501 (2005)

When a liquid flows over a surface, the traditional model says that the molecules of the liquid that are closest to the surface stay still, rather than sliding past. But experiments have suggested that surfaces are more slippery than this, particularly if the forces between the surface and liquid are weak.

The lingering confusion has been cleared up by Liliane Leger and colleagues at the Physics of Organized Fluids Laboratory, Paris. They used an optical technique to measure the velocity of two organic liquids, squalane and hexadecane, flowing over different surfaces. Not only do the liquids 'slip' even when the solid-liquid interactions are strong, but the amount of slip is also found to depend on molecular shape: branched molecules slip less.

S. HOPKIN/ARDEA.COM

JOURNAL CLUB**Dian J. Seidel**

National Oceanic and Atmospheric Administration, Silver Spring, Maryland

Follow the ups and downs of a boundary in the atmosphere with a climate scientist.

In introductory meteorology courses, we learn that the tropopause is the boundary between two layers of the atmosphere: the troposphere, where weather occurs, and the stratosphere above. I was

once content to think that the tropopause had little more geophysical importance than political borders on land. Not any more.

A recent series of path-breaking papers propose that the global tropopause is a sensitive indicator of human-induced climate change. From its first airing in the *Journal of Geophysical Research* (B. Santer *et al.* **108**, 4002–4024; 2003), this tantalizing idea sparked the interest of the broader climate-science community, and rekindled my own

research in the tropopause.

We expect the troposphere to warm and the stratosphere to cool in response to increases in greenhouse gases. And simple thermodynamic principles predict that the tropopause will rise. In support, Ben Santer and his colleagues detected an upward trend in the height of the tropopause in model-generated 'reanalyses' of atmospheric observations spanning the past few decades.

Motivated by Santer's work, and in collaboration with Bill Randel of

the National Center for Atmospheric Research in Boulder, Colorado, I am using data from weather balloons called radiosondes to study the structure and variability of the tropopause, particularly in extratropical regions. In winter, near the jet streams, for example, there can be multiple tropopause levels. Before we confidently rely on the tropopause to tell us about the climate, we need to understand its complexity and behaviour. Fortunately, we're making progress on this front.

NEWS

Japan consoled with contracts as France snares fusion project

PARIS

France has been chosen to host a billion-dollar international thermonuclear experiment. After an 18-month stalemate over whether Japan or France should host the project ITER, the Japanese finally bowed out on 28 June — in return for a hefty compensation package.

Scientists hope that, now the experiment's location has been decided, it could be up and running by 2015 (see 'Securing the funds'). If all goes well, it will be the first fusion experiment that generates more energy than it uses.

The promise of fusion is well known. Using the same reactions that power the stars, hydrogen nuclei can be fused to produce helium, releasing huge amounts of energy — and no high-level radioactive waste. But the line that usable fusion power is 40 years away, and always will be, is sadly just as familiar.

Recent progress has been promising, however, especially for tokamak reactors, in which hot plasma is confined in a floating doughnut shape by superconducting magnets. Both Europe's JET and Japan's JT-60 tokamaks have achieved short periods in which the energy released approaches the energy put in: JET holds the record, with a maximum power

output of 16 megawatts. At more than 12 metres across, the plasma ring in ITER will be about twice as big as JET's, and will hopefully generate 400–700 megawatts of power.

Negotiations over ITER's home have been deadlocked since December 2003. The United States and South Korea backed a Japanese site at Rokkasho, while China and Russia supported the European Union's bid for Cadarache in southern France. But at Tuesday's meeting of ITER's six international partners in Moscow, ministers finally agreed.

It's a deal

The European Union will now pay half of ITER's US\$5.5-billion construction costs, much of it coming from France. The other five partners will contribute 10% each, mostly in the form of equipment and components. Japan will win 20% of the manufacturing orders despite its 10% share. The European Union has also agreed to support a Japanese candidate for ITER's director-general, and Japan will provide 20% of the project's scientists, instead of the 10% to which it is entitled.

Up to 8% of the ITER construction budget will go towards partner facilities. These will

now be built in Japan, and three likely candidates are a supercomputing centre, an upgrade of the JT-60 and a materials testing facility.

Such centres would carry out research in parallel with ITER, so that if it is successful, work can start straight away on the next step — a prototype fusion reactor called DEMO. For example, the testing facility would use accelerators to find building materials that will stand the extreme conditions of a fusion power plant. Some even say that DEMO could be operational by 2030.

The Europeans are, of course, ecstatic. "There is no equivalent to this site anywhere in the world," says Jean Jacquinot, a plasma physicist who played a leading role in the Cadarache bid.

Cadarache is already home to more than 2,000 nuclear engineers and scientists, he points out. Most of them work on fission energy, but Jacquinot says ITER could still use their expertise. The site's use as a nuclear research facility means the project will have the necessary access to large amounts of electricity and water-cooling facilities, Jacquinot adds.

Meanwhile, Japan's researchers are looking for someone to blame. "Japanese scientists

Clear skies raise global-warming estimates

MUNICH

For more than a century, dust and aerosols in the atmosphere have been blocking some of the Sun's radiation, shielding us from the worst effects of global warming. The question has always been: how much? Now, as cuts in pollution allow the skies to clear, an attempt to quantify the effect on future temperatures has produced an alarming conclusion.

Even under relatively cautious assumptions about past and present aerosol cooling, the study suggests that global warming could easily exceed the upper extreme predicted by the Intergovernmental Panel on Climate Change (IPCC), as clean-air measures take effect.

"Things could get really uncomfortable," says lead author Meinrat Andreae, an atmospheric researcher at the Max Planck Institute of Chemistry in Mainz, Germany. "The climate system is much more sensitive to human per-



How much has pollution shielded us from the Sun?

turbations than has been thought. If our model is right, things could become totally uncontrollable in the second half of the century."

That is quite a big 'if', however. As cars, industries and power plants worldwide become cleaner, atmospheric concentrations of emitted aerosols are expected to drop substantially. But

how global temperatures will respond depends on how big the masking effect was in the first place — and that is the wild card in the climate game.

The problem is that different methods of estimating the cooling effect arrive at vastly different values. Trying to work it out from our understanding of how aerosol particles behave in the atmosphere suggests that the amount of solar energy reaching the ground will be reduced by anything from 0 to 4.5 watts per square metre. Working it out from a best guess of how sensitive the atmosphere is to greenhouse gases and how much warming we have seen so far gives 2 watts per square metre.

Such uncertainty has deterred researchers from estimating the effect of losing our aerosol shield. But as the skies are already starting to brighten (see *Nature* 435, 135; 2005), the question has become critical.



HYDROGEN CARS WILL SAVE LIVES

Cleaner vehicles would be better for people, as well as the planet.

www.nature.com/news

Securing the funds

The international fusion experiment ITER finally has a home. But don't assume there's a smooth ride ahead. Before a formal agreement is signed and construction can begin, ITER members must win the money that they have pledged from their respective governments.

Scientists hope this will happen by the end of the year, but several obstacles must be overcome. "Just because a site has been selected doesn't mean we're finished," says Gerald Navratil of the US ITER team.

The European Commission has committed money for 2006, for example, but it will need to double that in the budgets for 2007-13. There's no doubt that the commission sees ITER as a priority, but as budget negotiations for Europe's Framework 7 research programme have stalled, the way ahead is less than clear.

The situation is more hopeful in Japan, where officials say they can easily come up with the desired funds. But in the United States legislation could delay participation in ITER by almost a year: the House of Representatives passed an amendment to that effect that could become law. And in Russia, reports say the trade ministry has proposed that no new money go towards the fusion project in 2006.

After 20 years of fraught negotiations, it is crucial that governments agree on funding fast in order to keep political momentum behind the project, says David Baldwin, director of fusion programmes at General Atomics in San Diego. "Otherwise, the negotiators will die of old age."

G.B.

Victorious: Cadarache, in southern France, has been selected to host the ITER fusion project.

think it very regrettable," says one senior fusion scientist, who asked not to be named. He says one theory about why France won is that the choice of Rokkasho, which is fairly isolated, gave the impression that Japan was not serious about the project.

Others blame Prime Minister Junichiro Koizumi for being unable to match his French counterpart's charm. "Jacques Chirac could get up and talk for 30 minutes about the value of ITER and its significance for the future," *Nature* was told. "Koizumi would just throw

out one sentence: 'We really want to get ITER'."

Representatives from China and South Korea are said to be irked by the fact that Japan is getting so many perks despite their equal contributions. But the Americans just seem happy to have a decision. "I think the response from the community is one of relief," says Gerald Navratil, a plasma physicist from New York's Columbia University.

Declan Butler

Additional reporting by David Cyranoski in Tokyo and Geoff Brumfiel in Washington DC.

Andreae, along with German and British colleagues, used a climate model specifically designed to simulate aerosol effects. After calibrating it against a series of more complex global models, they plugged in a range of values for aerosol cooling, and ran the model to simulate future temperatures as the skies clear (see pages 1187-1190).

For a present-day cooling of 1.5 watts per square metre, which most climatologists agree is a relatively conservative value, the model implies that temperatures could rise between 6 and 10 °C by 2100. That is well in excess of the current IPCC predictions, which suggest a temperature rise of between 1.4 and 5.8 °C over the same time period.

Andreae acknowledges that there are many uncertainties about his study. But he points out that it is the best estimate we have so far. "This forces us to accept that pessimistic climate scenarios are much more

plausible than had been thought," he says.

Other experts are more cautious. "Climate modellers like playing around with values," says Theodore Anderson, an atmospheric scientist and aerosol researcher at the University of Washington in Seattle. "It is legitimate to engage in speculative reasoning. But I object to conclusions based on the assumption that our knowledge is better than it actually is."

Anderson points out that, for high values of aerosol cooling, Andreae's model breaks down, predicting unrealistically high or infinite temperature rises. He says this could mean that our understanding of what is driving the climate system is wrong. Or, he suggests, natural climate variability might be much larger than most scientists assume.

"The predictions may look more dramatic than what we actually expect," agrees Olivier Boucher, head of climate chemistry and ecosystems at the Hadley Centre for Climate

Prediction and Research in Exeter, UK. "But this is still an alarming hint at the upper bound of what can happen."

All agree that precise observations about the vertical distribution of aerosols are required. Calypso, a satellite funded by the United States and France, will provide such data after its launch in August, although scientists warn that it could take 20 years to get a clear picture.

In the meantime, Andreae says he hopes his results will rouse political debate, especially as the G8 summit looms. "Mankind must fight CO₂ emissions more aggressively," he says.

The uncertainty surrounding the effects of global warming has been widely used to imply that things might not be as bad as projected, says Michael Grubb, an expert on policy responses to climate change at Imperial College, London. "This study is a timely reminder that uncertainty also means things could be a lot worse," he points out. "Politically, this is a hugely important message."

Quirin Schiermeier

"Pessimistic climate scenarios are now much more plausible."

ON THE RECORD

“Stranger things have happened and craft have been found in unplanned orbits, so we are still hoping.”

Louis Friedman, director of the Planetary Society, talks about the disappearance of the solar-sail experiment, Cosmos 1, shortly after its launch on a Russian rocket.

“The Russian navy is searching for the debris.”

Vyacheslav Davidenko, spokesman for Russia's space agency, offers his own assessment of Cosmos 1's fate.

SCORECARD

**Learned debate**

A month after *The Lancet's* scathing attack on the Royal Society, the medical journal graciously adds that it would not go so far as to call the society a “stuffy, useless beast”.

**Macho ventures**

A French–Japanese consortium wants to build a jetliner that would travel at mach 5.5 — allowing scientists to zip to their meetings faster than a speeding bullet, literally.

**Political science**

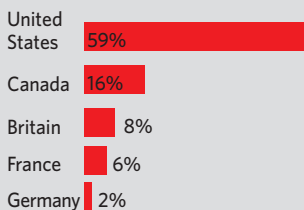
The American Civil Liberties Union accuses President Bush of a sustained assault on scientific inquiry, but officials retort that its study lacks credibility. The same might go for White House climate reports.

**Warm feeling**

Doctors deliver the first baby definitely created from frozen ovarian tissue — a ray of hope for chemotherapy patients.

DATAPOINT

Percentage of citizens who think the United States can be trusted to protect the global environment:



Source: Pew Global Attitudes Survey

Japan's university shake-up wins faint praise after first year

TOKYO

Last year, the Japanese government drastically reformed the country's university system. It stripped researchers of fixed benefits, but increased their freedom and independence — in the hope of making science more competitive. The government has just released the first survey on what researchers think of the changes, and the results suggest that they are unimpressed.

Before April 2004, the activities of national universities were tightly controlled by the education ministry, but scientists had secure funding, with guaranteed employment for life. Then, in the biggest reforms since the Second World War, universities were given semi-autonomous status. Researchers are now freer to collaborate with industry, and university presidents have much more freedom to decide how to spend their budgets. But institutes and researchers alike must earn their money — competing to win research grants on merit.

Reformers hoped that the change would reward promising young scientists in a system that has been crowded with unproductive professors (*Nature* 429, 207–221; 2005). But has it worked? To find out, the education ministry polled 2,000 researchers at universities, and at the public research institutes that have been going through similar reforms since 2001.

A spokesman at the education ministry

describes the results as “mixed”. On the positive side, 44% of the respondents say they are more willing to present their work outside their own labs in order to build closer relations with the public. And about 30% say they find it easier to work with private companies.

But many doubt whether the changes will raise overall standards in Japanese research. Only about 10% agree that research support is better or more flexible, and that research facilities have been improved by the changes. The fact that researchers have to compete for research grants has also not gone down particularly well.

About half of the respondents say that projects that meet specific industrial needs or lie in the government's strategic areas are more likely to be given money. Although this is what reformers were hoping to achieve, it leads many to express concern that important research fields are being ignored.

Kotoku Kurachi, a molecular biologist at the National Institute of Advanced Industrial Science and Technology in Tsukuba, believes that the reforms will help Japan to compete with countries such as the United States. But he says that unless the country's opaque grant system is overhauled, the changes will be in vain. “If we don't establish fair rules, it could get worse than ever.”

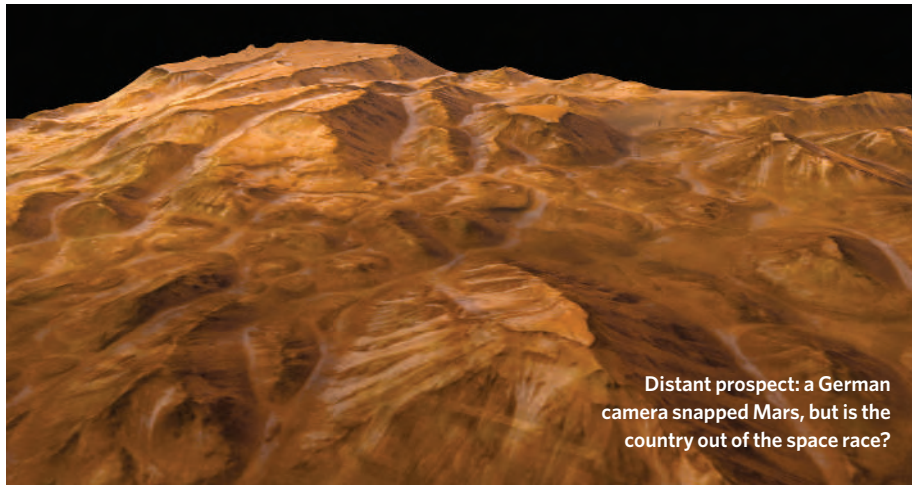
Ichiko Fuyuno

IMAGE
UNAVAILABLE
FOR COPYRIGHT
REASONS

Culture shock: Japanese universities' switch to semi-autonomy has provoked a mixed response.

T. WAGNER/CORBIS SABA

ESA/DLR/FU BERLIN/G. NEUKUM



Distant prospect: a German camera snapped Mars, but is the country out of the space race?

Shrinking budget grounds German space research

MUNICH

Germany was once a leader in space research. The Mars Express pictures that fired the world's imagination last year were taken by a German stereocamera, for example. And the APX 'sniffer' that first analysed the chemistry of martian rocks on NASA's 1997 Sojourner mission was also German.

But Germany's space research budget has shrunk by so much that the country's scientists say they no longer have the opportunity to participate in space missions. Frustrated by the government's failure to reverse the trend, they are warning that, without involvement in missions, the country's space-related research laboratories will inevitably wind down too.

Since reunification in 1990 put pressure on the country's spending as a whole, the space research budget has slipped every year. Its value is now only a third of what it was 15 years ago. Top astronomers, planetary scientists and climatologists have been discussing the crisis with politicians for months — but to no avail, say the researchers.

So on 21 June, the scientists published what they refer to as a "begging letter". No national space-science mission is planned for the next five years, they point out. And to make matters worse, they say they are unable to compete in placing instruments on international missions either.

Although Germany's space budget is shrinking each year, the country is locked into paying the highest contribution to the European Space Agency's science budget, based on

its high GDP. This means that Germany is paying for ESA's spacecraft and launches, but there isn't enough national money left over for German scientists to develop their own instruments to send on board, according to Günter Hasinger, a director of the Max Planck Institute for Extraterrestrial Physics in Garching.

For example, although the APX spectrometer is to be included in NASA's Mars Science Laboratory (due for launch in 2009 or 2011), it will be developed in Canada, not Germany — because the department that first worked on it, at the Max Planck Institute for Chemistry in Mainz, has closed.

The extraterrestrial-physics institute is also rethinking its strategy, and has abandoned gamma-ray astronomy to concentrate all its efforts on X-ray astronomy, says Hasinger.

And environmental scientists are complaining that they don't have enough money to exploit Europe's Earth observation satellites. "Germany is a big contributor to Earth observation hardware," says Jochem Marotzke, an oceanographer and a director at the Max Planck Institute for Meteorology in Hamburg. "But we have little project money to make use of it."

When asked about the crisis, a research ministry spokesman told *Nature* that space research remains a "high priority" for the German government. Scientists are hoping that, with an election looming, their public plea will help persuade politicians to back that sentiment with further funds.

Alison Abbott

SPECIAL REPORT: SCIENCE & AFRICA

IMAGE
UNAVAILABLE
FOR COPYRIGHT
REASONS

C. HUGHES/PANOS

A message to the G8 summit

When the G8 leaders meet in Scotland next week to discuss how to help Africa's poorest nations, they are unlikely to hear the chants of the protestors — an 8-kilometre fence around their luxury hotel will see to that. But the activists have, to some extent, already been listened to: a debt-relief package has been signed by the group of eight industrialized countries and a hike in aid is also on the cards. But when it comes to spending this extra money, one question is whether the voices of Africa's scientists will be heeded.

On the following three pages, *Nature* presents those voices. They need to be heard, as science and technology are more of a priority for aid agencies than ever before. African universities, for example, are the subject of a new focus by the World Bank. Africa's leaders have also singled out science and technology in their continent-wide political strategy — the New Partnership for Africa's Development (NEPAD).

The comments that follow make for challenging reading. Every area seems to require immediate attention, from disease and climate change to a lack of access to education and sanitation. But themes emerge nonetheless. Solutions must factor in the needs of local communities and environments. Projects should be run as far as possible by Africans, not the donors. And Africa needs long-term backing from rich nations, not an uncertain future in which aid waxes and wanes. If science and technology projects are to help shape Africa, these are the strategies that should shape them.

Interviews by Peter Aldhous, Declan Butler, Jim Giles, Michael Hopkin, Mark Peplow and Quirin Schiermeier



NOAA/NASA

SOUTH AFRICA

John Mugabe

Adviser on science and technology to NEPAD, based in Pretoria. Helped to establish the partnership's African Forum on Science and Technology for Development (AFSTD).

I spend most of my time working with governments and donors to ensure that scientific knowledge is incorporated into African skill sets, policies and strategies. We need more capacity for African countries to apply science to their problems, focusing on health, water, agriculture and the environment, and to generally increase economic competitiveness.

A big part of this will be technological innovation. You can never say when you have 'enough' technologies. There are many technologies available to manage water supply, for example, but few to improve water quality.

The hope of every country in the world is to have more scientists. But for us in Africa it is difficult. There is not just the task of training more scientists, we also need to create solid institutions and ensure that our scientists have specific, well-resourced projects to work on.

Debt relief will help, but developed countries need to ensure that the money is going to benefit productivity in Africa. Knowledge needs to be shared better between developed countries and Africa, to enable African countries to improve their technologies.

But in the short term, we want to see a commitment from the G8 to put together an African science fund that would be available to countries on a fairly flexible basis to address their problems — not necessarily without certain minimum conditions. That will be a better way for African scientists to get the relatively small amounts of money they need to work on projects that will benefit Africa.

KENYA

Kevin Marsh

Epidemiologist and director of the Wellcome Research Programme at the Kenya Medical Research Institute, Kilifi.

Despite receiving 30 years' worth of aid to develop medical science, Africa simply hasn't managed to build up a high-quality research network. Outside South Africa, there are probably fewer than a dozen health researchers with a high international profile, actively driving big, imaginative programmes that draw in substantial funding and scientific interest from around the world.

Nothing will change until research initiatives start coming from within African research centres. Without an established career structure, it is hard to get that critical mass of expertise.

Because there is no research culture at many African universities, people do not see science as an attractive option. And once trained, the best African scientists are often attracted abroad. If you want good scientists in Africa, you need to pay them.

A continent-wide group is putting together a detailed plan for developing medical science in Africa. We hope to discuss these proposals with the UK government in the next few weeks. We need to develop training programmes and set up collaborative research links across Africa and abroad, rooted in African health problems. We can start by focusing on the African research centres that are already doing world-class research. We need to turn them into engines for training home-grown PhDs and postdocs, and they can also play a critical role as the nodes of an African research network.

The G8 nations should act on the Commission for Africa report drawn up in March, committing up to US\$3 billion over 10 years for African centres of excellence in science and technology. Of that funding, US\$900 million should go towards developing a vibrant research network that links these centres.



RWANDA

Romain Murenzi

Rwandan minister of education, science, technology and scientific research. Previously a professor of theoretical physics at Clark Atlanta University in the United States.

Ten years after its gruesome civil war, Rwanda is still widely associated with the second-worst genocide of the twentieth century: the conflict between the Hutus and the Tutsis that killed 800,000 people in just 100 days. Less widely known are Rwanda's serious attempts to promote science, technology and education as a means to combat poverty, backwardness and conflict.

Since the war, the number of people being trained in Rwanda to degree level in science has increased eightfold. And 25,000 students are enrolled at the National University of Rwanda and other higher-education institutes. By 2020, we hope that about 100,000 Rwandans — around 1% of the rapidly rising population — will have a higher academic

GHANA

Fred Binka

Head of epidemiology at the School of Public Health, University of Ghana, and executive director of the INDEPTH pan-African network of field epidemiological centres.

If the G8 summit is really to improve the lot of Africa, it must make a big commitment to controlling infectious diseases, particularly the three big killers: AIDS, malaria and tuberculosis. Without this, all other efforts to raise Africa out of poverty will be futile.

That is because, although it's accepted that poverty causes disease, it is too often overlooked that disease causes poverty. AIDS is destroying African economies and workforces, and malaria, my own field of work, is estimated to slow African economies by 1.3% a year — crippling the poorest households and workers.

Plans for debt relief, easing of trade barriers and new development funds will be good news for Africa. But these measures address just one side of the problem — poverty. If we do not also tackle health we will just continue in a vicious circle, where disease breeds poverty, which breeds disease.

"Without controlling infectious diseases, all other efforts to raise Africa out of poverty will be futile."

The Commission for Africa report mentions health research, but it is vague. The fact is that malaria is right here to be controlled. The situation is getting better, and more money is flowing in. But even the Global Fund to fight AIDS, Tuberculosis and Malaria is still supporting fragmented efforts; it is not taking the bull by the horns.

What the G8 needs to advocate is much wider use of current intervention tools. We need to be putting 50 million bednets a year into Africa and we need to scale up drugs and house-spraying to cover entire countries. Zambia recently pledged to do this and aims to cut mortality by 75% in three years. When I first heard this I was really excited — I'm not joking; if the G8 backs and helps fund that approach across the continent, we will reduce the malaria burden dramatically.

To create major change we need African-led efforts, supported by the international community. I think the message is also getting through to the G8 that we need research, and that we cannot do this without scientists, trained and working in Africa. The G8 must make a strong commitment to building human capacity, and this must not be lip-service.

SPECIAL REPORT: SCIENCE & AFRICA

SOUTH AFRICA

Mark Henning

Chair of South Africa's Sector Education and Training Authority.

Education is vital. All sorts of studies say that the best thing African countries can do is to have universal primary education, including bringing girls into schools. That should be a priority. It's always been assumed in Africa that if you get a university degree your future is secure, but if you don't, then life holds nothing. That's false. We need to find a balance between providing high-quality

"It's got to be the right education for the right people."

specialized education and uplifting the poorly educated majority.

The relationship between education and economic growth is complex. Economic success depends on many things other than education, and it has to be the right education for

the right people. Zimbabweans are well educated, for example, but the country's disastrous economic policies mean that people are starving.

Malaria and HIV/AIDS are huge problems too. Where you have family groups headed by ten-



year-old children, it has a profound effect on education. Financial support and development of drugs by the G8 can make a big difference.

Youth unemployment is a time bomb — well over 60% in South Africa. 'Trade not aid' is a critical slogan. The G8 needs to free up trade in a way that will let African nations stop relying on aid. With 60 undernourished children in a class it's hard to make progress. And don't forget the cancer of corruption — a lot of that is coming from G8 countries, whose entrepreneurs are buying favours. Fix these problems, and maybe educational reforms can deliver.

qualification. And attempts to interest Rwanda's young generation in science will start early. From next year, all the country's 2,200 primary schools will be equipped with 'science corners', displaying basic information about the Sun and the planets, the cycle of life, or a map of the world — plus a computer with an Internet connection.

Scientists and engineers will find plenty to tackle in Rwanda, from soil erosion to water management, health, biodiversity and ecosystem conservation. But one particular problem is that there are not enough science teachers, as many were denounced by their colleagues and killed during the civil war.

The horrible conflict was partly rooted in academic circles: some intellectuals sowed the ethnic hatred that led to the genocide. It is now crucial that Rwanda creates an education system that rewards merit, rather than ethnicity. Such a system can become a model for the reconciliation of Rwandan society.

To this end, an ambitious national science, technology and innovation project was launched at a conference in May, supported by science-policy experts from Australia, Britain, Sweden and the United States. It aims to give children and young people access to basic and higher education, to strengthen

"Rwanda must create an education system that rewards merit, rather than ethnicity."

human rights and peace education, and to eventually transform the country into a knowledge-based economy. The Rwandan government has requested US\$130 million from the African Development Bank for the programme. Additional support from the G8 countries is both necessary and very welcome.



MOZAMBIQUE

Pascoal Mocumbi

Prime minister of Mozambique from 1994 to February 2004. Now high representative of the European & Developing Countries Clinical Trials Partnership programme.

Decisions taken at the G8 summit will be important in building the strong science and technology base that Africa needs, particularly for developing the tools needed

to control disease. The major killers are causing devastation in Africa. In my own country, Mozambique, one in four children dies before the age of five, and malaria is the biggest cause. What is lacking at the international level is an understanding of the importance of research and development, particularly within

Africa, in the fight against disease.

Much can be done to fight malaria with existing tools. But we need more effective drugs, and we need a vaccine. We need research and clinical trials in Africa. On vaccines there is light at the end of the tunnel — the number of candidate vaccines is growing. But the G8 needs to push for a more coordinated effort. We need a new international malaria vaccine enterprise, drawing, for example, on the work of the Malaria Vaccine Technology Roadmap, a project in which I am involved.

The G8 must also address infrastructure, particularly for health research. Things are much better than they were ten years ago. African countries now have their own draft strategy for science and technology, drawn up by the AFSTD (established by NEPAD) to promote their economies and reduce poverty. The G8 must build on this to achieve sustained progress. Debt relief is an important decision, and African countries must plough some

of this money into science, technology and health, where it could make a big difference.

Education is a key plank: we need nothing less than a comprehensive system from primary school to higher education. Another is telecommunications. The Internet is already revolutionizing African research, but bandwidth remains too slow and expensive.

But most important, the G8 must urgently respond to the call from Africans for a flow of predictable funding to support NEPAD's strategy. We need to think at the country level; change will only come when the leadership of the country is in the driving seat. Take the UN Millennium goals to slash poverty and diseases. Unless these are tackled at the country level, we will simply continue to have more talks, more meetings, and little progress.

NIGERIA

Anthony Nyong

Expert on environmental resources and natural hazards at the University of Jos.

Poverty is a major cause of environmental degradation and causes people to live unsustainably. Take deforestation: people who cut down trees don't do it for fun: it is a bid to survive. Much of the rural population depends on wood as fuel for domestic energy and cooking.

Faced with the need to survive, people even have to encroach on protected forests and game reserves. It is unfair and impractical to think that force can prevent this. Africans need appropriate science and technology to develop cheap and affordable energy sources.

Climate change is likely to make matters worse: major international reports conclude that Africa is the most vulnerable continent. A first step towards reducing this vulnerability is to assess the potential impacts of climate change. But most African nations have neither the capacity nor the technical ability to do this. The few studies that do look at Africa have



largely been conducted by Western scientists. Africans need to build scientific capacity so that we can develop our own models, validated over Africa.

Africans also need science and technology to help adapt to predicted changes — to develop affordable, accessible and sustainable tools, such as early warning systems, drought-resistant crops, water-extracting and harvesting systems, and flood-protection.

It is time to stop the 'mercenary' form of development that has long been practised. Africa does not need food aid that continually impoverishes its own people. We need to enable farmers to grow their own food in the face of environmental challenges.

SOUTH AFRICA

Mike Jensen

Telecommunications consultant working with NEPAD and international agencies, and an expert on African Internet connectivity.

Improving Internet connections for research centres and hospitals in Africa would be one of the most cost-effective actions the G8 could take. It would empower African scientists and medical researchers by giving them high-speed access to the wealth of information and vast scientific databases available on the Web.

Although some African countries have high-speed Internet connections, Africa lags far behind the rest of the world. The root problem is that telecoms in Africa are national monopolies. High-speed Internet connections usually have to go through the national operator, and bandwidth costs 10 to 100 times more than in Europe or North America.

The G8 should encourage liberalization of these markets. Meanwhile, the international community should subsidize national research networks directly. The European Union has already connected North African countries to Europe's research network. And the provision of high-speed satellite links to malaria research centres has shown how the Internet can boost productivity and cooperation.

Research centres and networks need to band together to increase their bargaining power in negotiations over bandwidth costs. The G8 should support such initiatives, including NEPAD's plan to create a network of optical fibre cables across Africa. It should support the Indian government's project to fund a satellite for educational purposes for the African Union. And it could address the lack of computing infrastructure and human resources in scientific research centres.

But we have to start somewhere: the creation of Internet-connected research centres in universities and hospitals would be a good place.

KENYA

Florence Wambugu

CEO of A Harvest Biotech Foundation International, a Kenyan organization dedicated to promoting sustainable agriculture through the use of biotechnology.

We cannot develop Africa without biotechnology. Enormous numbers of people suffer from malnutrition in some regions, and this is where biotechnology has huge potential.

One example is NERICA (New Rice for Africa), a variety developed by the West Africa Rice Development Association in Bouaké, Ivory Coast. The rice was created by conventional breeding and combines high-yield Asian strains with drought-resistant African ones. It is a good example of the research and development we can do when there is partnership between scientists in Africa and abroad.

But we have to take a holistic approach — we also need to address other issues such as soil fertility, water management, human infrastructure and capacity development.

The problem is that there is a disconnect between high-level international research and the perspectives and priorities of African leaders. Most research here is donor-funded. There is an urgent need for African countries to fund their own research so that they have

a stake in the results. That way the results will be more relevant and can be linked to local communities.

Involving rural people is crucial. The poverty in Africa is in the villages. We need education and training for farmers so that they can make use of opportunities such as improved seed banks. That will empower them. You can't just give them an agricultural innovation and leave them to it. I believe in science and technology, but the way it is implemented is very important.

For example, genetically modified (GM)

crops have a major role to play in Africa, especially in tackling problems such as pests, drought and malnutrition. To succeed, GM technology must be implemented in a way that gives Africans true ownership. Although there is room for many different players, including the private sector,

researchers and agricultural organizations, greater emphasis should be placed on collaborations with countries outside Africa. When it comes to staple crops, the possibility of royalty-free technologies must also be explored.



SIERRA LEONE

Ogunlade Davidson

Professor of mechanical engineering and energy systems at the University of Sierra Leone in Freetown, and co-chair of Working Group III of the Intergovernmental Panel on Climate Change.

Development in Africa will need energy. A Latin-American household consumes 3.5 times more energy than a sub-Saharan one, and North Americans consume 22 times as much energy as we do. Almost 75% of energy in Africa is consumed by South Africa and six North African countries — the other 46 countries account for a mere quarter of consumption. Sub-Saharan countries will demand substantial amounts of extra energy in the next decades.

In theory, Africa has abundant energy resources — oil, gas, coal and hydropower, depending on the region — to meet the increased demand. But enough energy does not reach consumers, and the quality of refineries and fuel products is often poor. Two thirds of the gas obtained as a by-product of

oil production is wasted, because producers claim there is no demand for it. Here is a case for the G8 countries to put pressure on the big oil companies, because African governments are too weak to urge them to invest in new and more energy-efficient infrastructure and in power systems that use gas.

Climate change is also a key issue, because although Africa contributes least to greenhouse gas concentrations, it is most vulnerable to global warming. Malaria, sea-level rise, droughts and other impacts on rain-fed agriculture are some of the most difficult problems facing the continent. Rainfall has decreased substantially since 1900, which can partly be

attributed to global warming. And air pollution from antiquated coal plants and cooking fires is a widespread problem in countries such as Zimbabwe and South Africa.

For all these reasons, it is essential that Africa's growing hunger for energy is satisfied by clean and environmentally friendly means. The rich countries should help by promoting advanced production of renewable energy, such as from biomass.

"The G8 countries must put pressure on the big oil companies to invest in new and more energy-efficient infrastructure."

Brazil's biosecurity law faces legal challenge

Brazil's rosy future as a biotechnology haven is in doubt after claims that a biosecurity law contravenes the country's constitution. Cláudio Fonteles, the attorney-general, has asked Brazil's Supreme Court to consider whether the four-month-old law is unconstitutional (see *Nature* 434, 128; 2005). The law legalized research using human embryonic stem cells and established a system for the approval of genetically modified crops.

On 21 June, Fonteles challenged a section of the law that assigns the decision about whether a genetically modified organism is environmentally safe to a special committee attached to the science ministry, and not to federal and local governments.

Fonteles also argued that the section of the law that legalizes stem-cell research is unconstitutional, as the constitution protects the right to life. He mustered the support of more than half a dozen scientists.

Fonteles is expected to leave office this week. But his successor, Antônio Fernando Souza, will almost certainly support the legal challenges.

IMAGE
UNAVAILABLE
FOR COPYRIGHT
REASONS

In the front line: Tamiflu would be the drug of choice in the event of an avian flu pandemic.

Flu drug's inventor acts to regain control from Roche

Companies are sparring over rights to the drug that is seen as the best hope for treating human cases in an avian flu pandemic.

Oseltamivir, branded as Tamiflu, is currently marketed by Swiss firm Roche. But last week the drug's inventor, Gilead Sciences, based in Foster City, California, said it wanted to end a 1996 agreement that gave Roche exclusive marketing rights. It accused Roche of failing to promote the drug sufficiently and of underpaying royalties. If the companies fail to agree a settlement within 90 days, the dispute will go to arbitration, which could take 18 months.

Public-health experts say they do not expect the dispute to disrupt governments' short-term plans to stockpile the drug.

Germany ends impasse over university funding

Ending an 18-month political deadlock, the German government last week finally approved a €1.9-billion (US\$2.3-billion) programme for strengthening university research. Under a deal agreed with Germany's 16 state governments, the country's main research agencies, including the Max Planck Society, are also guaranteed annual budget increases of 3% until 2010.

Plans to create a number of elite universities had been first announced in January 2004. But the plan was threatened by disagreement between some states and the federal government over who was responsible for science and education (see *Nature* 433, 448; 2005).

Under the revised programme, which runs from 2006 to 2011, competition for creating ten elite universities will be supported with €13.5 million per year. But the bulk of the €1.9 billion is earmarked for the creation of 30 'clusters' involving several universities. A total of €40 million

per year will be spent on postgraduate programmes.

Speaking at the Max Planck Society's annual assembly in Rostock last week, German Chancellor Gerhard Schröder promised to make science a high priority if he wins the election in the autumn. He said he would make an extra €15 billion available for science and education by 2010.

Japan targets more whales despite losing key votes

Conservationists won a hollow victory last week in Ulsan, South Korea, at the International Whaling Commission's annual meeting.

Japan failed in its bid to overturn a moratorium on commercial whaling, and was outvoted on several other initiatives, such as starting limited coastal whaling along Japanese shores and abolishing an Antarctic whale sanctuary. It also lost a vote to expand its 'research' whaling programme by doubling its take of minke whales and starting to take fin and humpback whales.

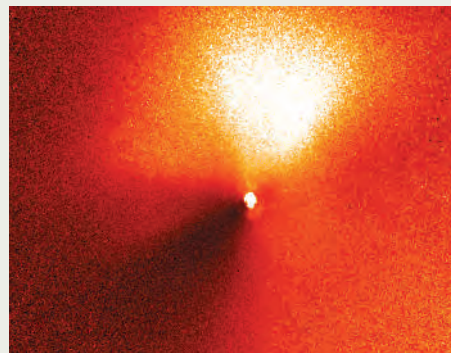
But the last vote has little more than political significance, as Japanese delegates told *Nature* that they will go ahead with their plan to increase the whale take anyway.

Hubble captures comet's dusty outburst

Comet Tempel 1 spits out a jet of dust in this image snapped by the Hubble Space Telescope on 14 June.

The comet's 7-kilometre-wide nucleus appears as a bright spot at the centre of the picture. Dust sprays out for 2,200 kilometres above it, in a jet that was not visible just seven hours earlier.

Astronomers are hoping for a similar outburst, which could unleash primordial material trapped inside the comet, after NASA's Deep Impact spacecraft deliberately slams into Tempel 1 on 4 July.



NASA/ESA/P. FELDMAN/H. WEAVER

Biologists weigh up risks of synthetic genomes

Leading figures in the emerging field of synthetic genomics are launching a study to investigate the risks and benefits of the technology. They hope to provide guidance to government officials who may in future step in to regulate the field.

Using state-of-the-art DNA synthesizers, scientists can now manufacture stretches of DNA up to a few tens of kilobases long. The technology can be used to alter organisms

more profoundly than conventional genetic engineering, and might one day be used to create a completely synthetic organism.

The Massachusetts Institute of Technology (MIT) and the J. Craig Venter Institute in Rockville, Maryland, are teaming up with the Center for Strategic and International Studies in Washington DC to launch a series of workshops to consider the implications of the technology — including the risk of it being used to create biological weapons.

"The development of synthetic genomics is potentially a double-edged sword," says Drew Endy, a synthetic biologist at MIT.

China's burning ambition

The economic miracle that is transforming the world's most populous nation is threatened by energy shortages and rising pollution. It also risks plunging the planet's climate into chaos. **Peter Aldhous** reports.

China is booming, and its hunger for energy is insatiable. For its people, the dismal air quality across much of the country is a constant reminder of its reliance on coal and other dirty fuels. When *Nature* visited Beijing to meet the technocrats responsible for China's energy policy, the city was blanketed in acrid smog. After just a few days of stagnant weather, visibility in some districts had dropped to tens of metres. Flights were delayed and the Beijing Environmental Protection Agency advised people to stay indoors. You could almost taste the sulphur in the air.

Energy and its consequences for health and the environment are high on the Chinese political agenda. But the hard-headed approach of the country's leaders should give us all pause for thought. China's energy policy will continue to be based around coal, they say, so the question of whether this notoriously filthy fuel can ever be made 'clean' is central to the country's development — and to the long-term stability of the global climate.

The most immediate problem for China is that its economic growth is already outstripping its energy supplies. In boomtowns from Shenzhen to Chengdu, electricity is now an unstable commodity. Last year, 24 of China's 31 provinces, municipalities and autonomous regions admitted that they lacked sufficient power. In the summer, when drought curtails

hydropower and air conditioners surge into life, blackouts have become commonplace.

The nation's coal mines are straining to meet the demand, at a terrible human cost. According to conservative official estimates, more than 6,000 workers were killed in China's mines last year — making them the world's most dangerous — and the death rate was undiminished in the first half of 2005.

Most coal-related fatalities never make the headlines, however. Many Chinese cities fail to meet international — or even their own — standards for air quality, causing hundreds of thousands of premature deaths each year. China's increasing use of coal is also sending CO₂ emissions skyrocketing, threatening a global climate disaster. "We understand that coal means not only energy, but also social and environmental impacts in the long term," says Zhou Enshi, director-general of the Energy Research Institute in Beijing and a leading proponent of the country's energy strategy to China's leaders.

While Dadi and other senior energy planners recognize these problems, their enthusiasm for coal remains strong. The country's leaders are determined that its economy will quadruple in size by 2020, which will require at least a doubling of the energy supply. Coal will bear most of the burden. "We have to increase coal consumption," says Guo Yuan, an energy systems analyst at Dadi's institute. "It's not a good picture, but we have to do it."

LI JIANGSONG/IMAGINECHINA

IMAGE
UNAVAILABLE
FOR COPYRIGHT
REASONS

Electricity generation is by far the biggest consumer of energy, although the demands of the transport sector are growing fast. Between 75% and 80% of China's electricity is generated by burning coal. Another 20% comes from large-scale hydropower projects, with most of the rest coming from nuclear stations. As yet, oil, natural gas and renewables such as wind barely feature in the electricity mix. But by 2020, according to official projections, gas-fired stations could be meeting 15% of China's electricity needs, while nuclear power may have expanded to around 5%. And thanks to a law passed in February this year designed to promote renewable energy, wind and other renewables could account for 10%. However, with power demands poised to double over the same period, it's clear that a massive increase in coal consumption is unavoidable.

Sustaining economic growth is the leadership's priority, say seasoned China watchers, but it wants to achieve this without compromising energy security. China lacks substantial reserves of oil and natural gas, and is determined not to become heavily dependent on imports. But the country has coal in abundance. So it will use the fuel in ever-larger quantities, mainly to avoid a reliance on Russian oil and gas that could eventually bring the two powers to the brink of war.

But can China meet its energy needs without poisoning its environment and filling the lungs of millions of people with particulates and oxides of sulphur and nitrogen? The effects of acid rain are spreading, and there are suggestions that soot is already disrupting the regional climate (see 'Brown clouds cast a dark shadow', overleaf).

Global climate change doesn't yet loom large in the thinking of China's leaders, but international experts note with alarm that coal is the worst offender in terms of CO₂ emissions. "The global problem is climate. But for China, conventional pollution is the main problem," says Li Zheng, who directs the Tsinghua-BP Clean Energy Research and Education Centre, a collaboration between Beijing's leading scientific university and the British energy firm.

Efficiency drive

China's energy planners have realized that improving energy efficiency is the easiest way to promote economic growth while controlling pollution. "China should work first on this," says Dadi. Predictions that assume 'business-as-usual' suggest that total energy demands will rise to the equivalent of 3.5 billion tonnes of coal per year by 2020. But introducing a suite of measures to improve efficiency could keep that below 3 billion tonnes, says Dadi. "Technically, it's do-able."

This new drive for efficiency stems in part from a quietly influential initiative run by the San Francisco-based Energy Foundation. Bankrolled for a total of US\$40 million since 1999 by the Hewlett and Packard foundations, the China Sustainable Energy Program is

GONG CHAI HIN/AP/GETTY IMAGES

IMAGE UNAVAILABLE FOR COPYRIGHT REASONS

Smog city: China's energy crisis is boosting interest in new technologies like coal liquefaction (below).



working with Chinese energy researchers to improve efficiency and cut pollution. Priorities include new efficiency standards for buildings, appliances and vehicles, and promoting renewable energy sources. Fuqiang Yang, who heads the Energy Foundation's Beijing office, points to recent successes such as the renewable energy law, plus fuel-efficiency standards and energy-efficient building codes adopted by central and local governments.

Energy efficiency is an admirable goal, but China's appetite for growth and the leadership's desire to limit imports of foreign oil mean it won't be enough. So China is embracing technologies that, in the West, remain on the fringes. Du Minghua, director of the Beijing Research Institute of Coal Chemistry, sees coal as an energy panacea, able to meet China's demands for electricity, liquid fuels and gas. "Coal is the solution for all three," he exclaims, before launching into a presentation on his institute's work on coal gasification and liquefaction.

Finding ways to reduce dependence on oil, critical for the transport sector, is the top priority for Minghua's institute. Young coals such as lignite can be converted straight to liquid fuels by heating them to 450 °C with hydrogen and a suitable catalyst, Minghua explains.

Older coals such as anthracite must first be heated in oxygen to produce a mix of hydrogen and carbon monoxide known as syngas, which can then be converted into liquid fuels. Some of these can be blended with diesel and pumped straight into a conventional engine.

Despite Western experts' scepticism about the direct coal-to-liquid technology¹, the state-owned Shenhua Group is now building the world's first commercial direct coal-liquefaction plant in Inner Mongolia, scheduled for completion by 2008. And China is also in discussions with the South African company Sasol about the possibility of building two large indirect liquefaction plants.

Crude substitute

Neither process is a model of efficiency, however. Direct liquefaction is about 60% energy-efficient, indirect techniques around 45%. But China's desire to seek alternative liquid fuels is so great that Minghua estimates that liquefaction technologies could be providing it with more than 50 million tonnes of fuel per year by 2020. "This is a personal estimate," he stresses — but one that will be music to the ears of China's leaders. If Minghua is correct, coal liquefaction could reduce China's demand for crude by 100 million tonnes per year, or about one-third of its anticipated imports by 2020.

Coal is also central to the thinking of researchers at the Tsinghua-BP centre. Zheng is focusing on a strategy called polygeneration in which a single plant would convert coal into syngas, then use it in gas turbines to generate electricity and also convert it into liquid fuels². Sulphur is removed as an integral part of gasification, cutting pollution. To demonstrate the technology's potential, Zheng and his colleagues have conducted a 'syngas city' simulation for Zaozhuang in the eastern Shandong Province. Like many industrial centres in

SHANXI COAL CHEMISTRY INSTITUTE

Brown clouds cast a dark shadow

China's flood season officially started this month with destructive floods in many parts of the country. In the past 20 years it has seen increasing summer floods in the south and drought in the north. The likely culprit is air pollution and, as this escalates with China's rapid industrial growth, it could alter weather across the region.

The key player in China's climate woes is the blanket of aerosol particles that hover over Asia. China isn't alone in creating this pollution hazard. India is a major contributor to the brown clouds of smog — mostly black carbon, organic carbon and other aerosols such as sulphates and nitrates — formed by wildfires and by burning fossil fuels and biofuels.

Black carbon, a sooty by-product of coal-burning, absorbs sunlight, resulting in a hotter atmosphere and cooler ground. Sooty particles also affect rainfall by seeding smaller droplets and

preventing the formation of larger droplets. This aids cloud formation, but reduces the amount of rain produced.

To simulate the observed changes in China's rainfall patterns in recent decades, a team led by Surabi Menon of the NASA Goddard Institute for Space Studies in New York used a global climate model that factored in black-carbon emissions⁴. But although climatologists generally agree that aerosol pollution has altered China's rainfall, they remain cautious about its potential regional impact.

"We are dealing with imperfect measurements and imperfect models," says George Carmichael of the University of Iowa. Reliable measurements of aerosol emissions are lacking, particularly for black carbon. And climate models are riddled with uncertainties, for example how aerosols modify clouds.

Even so, studies reveal a similar picture elsewhere.

IMAGE
UNAVAILABLE
FOR COPYRIGHT
REASONS

Simulations by Veerabhadran Ramanathan from the Scripps Institution of Oceanography in La Jolla, California, and colleagues show that aerosol pollution caused changes over the north Indian Ocean that resulted in decreased monsoon rainfall and increased drought in India⁵. Similarly, China's pollution could affect surrounding oceans, altering monsoon rainfall across the region, says Ramanathan.

The next step is to reduce some of the uncertainties. Project Atmospheric Brown Clouds, run by the United Nations Environment Programme, began monitoring Asia's smog earlier this year. And improvements in satellite measurements of aerosols, together with China's plans to increase emission monitoring, will help determine the extent and impact of the country's air pollution. **Carina Dennis**

China, Zaozhuang faces a major problem: how to continue growing when the only readily available fuel is high-sulphur coal.

In the 'syngas city' model, the Zaozhuang authorities would provide incentives to promote polygeneration, which not only generates electricity but also produces methanol for vehicle fuel and dimethyl ether for domestic cooking and heating. The simulation suggests that polygeneration could meet more than a quarter of Zaozhuang's electricity needs by 2020. It would also achieve drastic cuts in sulphur dioxide emissions while reducing the need to invest in expensive flue-gas desulphurization technology at conventional power plants³. Further reductions in air pollutants, such as ozone-forming compounds, would come from the wider use of methanol and dimethyl ether.

Such simulations are the stock-in-trade of energy researchers worldwide. But in China there may be a greater chance of their being implemented, given the authorities' power to enforce their will. Preparations for the 2008 Beijing Olympics are a case in point. Realizing that the city's appalling air quality could damage athletes' health — and present a poor image of China to the world — the city is now engaged in a frantic clean-up, closing some 200 heavily polluting factories, piping in natural gas, and introducing a clean 'bus rapid transit' system. "The Olympics are a very big

opportunity," says Li Hao, who heads Earth-View, a Beijing-based environmental group.

Zheng and his colleagues hope that growing official concerns about environmental health will also boost their proposal to build a poly-generation demonstration plant, costing some 5 billion yuan (US\$600 million), which would generate up to 400 megawatts of electricity and produce as much as 400,000 tonnes of liquid fuel per year. "We got a very good response from the government," says Zheng.

Greenhouse city

But while polygeneration and other clean-coal technologies may help to scrub China's filthy air, they won't do much in the short term to limit the nation's growing greenhouse-gas emissions. According to Zheng's simulation, total CO₂ emissions from power plants would be higher for the syngas city than if Zaozhuang continues using conventional technologies³.

In the long run, however, polygeneration could provide a route to a more sustainable future, in which hydrogen is extracted from syngas and used to power fuel cells, while CO₂ is captured and sequestered. "But to get there, the investment will be huge," warns Zheng.

Given the costs involved, experts say that China's interest in carbon sequestration will depend largely on the willingness of Europe, North America and Japan to pay for it. Those who work in the energy industry are blunt

about China's determination to strike a hard bargain. If the necessary cash isn't forthcoming, they say, all deals are off.

China's potential to single-handedly emit enough CO₂ to negate all other nations' efforts to control their greenhouse-gas emissions could place its leaders in a strong negotiating position. "If it's business as usual, then the planet is dead," says David Moskowitz, director of the Regulatory Assistance Project, based in Gardiner, Maine, who is advising Chinese officials on reforming the electricity-generation sector.

That should provide food for thought for the leaders of the G8 wealthy nations, who meet in Scotland in July with global warming on their agenda. China is a signatory to the Kyoto Protocol on climate change, but as a developing country it doesn't yet have an emissions reduction target. Whatever strategy world leaders contrive to save the planet, China will sooner or later have to be brought on board. And that won't come cheap. ■

Peter Aldhous is *Nature's* chief news and features editor.

1. Williams, R. H. *et al. Energy Sustain. Dev.* **7**, 103–129 (2003).
2. Zheng, L. *et al. Energy Sustain. Dev.* **7**, 57–62 (2003).
3. Hongtao, Z. *et al. Energy Sustain. Dev.* **7**, 63–78 (2003).
4. Menon, S. *et al. Science* **297**, 2250–2253 (2002).
5. Ramanathan, V. *et al. Proc. Natl Acad. Sci. USA* **102**, 5326–5333 (2005).

For more on China's environmental problems see page 1179.



Big plans for little brains

Experts in neuroscience, computing and education are coming together in a massive effort to put the way in which children are taught on a sounder scientific footing. **Trisha Gura** profiles this ambitious — some might say foolhardy — initiative.

In 1997, John Bruer, the president of the James S. McDonnell Foundation, launched a broadside against the fashion of taking findings from neuroscience and trying to apply them in the classroom. Experienced in both cognitive science and education, Bruer set out to demolish the hype surrounding what he saw as blatant pseudoscience.

"Currently, we do not know enough about brain development and neural function to link that understanding directly, in any meaningful, defensible way, to instruction and educational practice," he wrote, arguing that so-called brain-based curricula had crossed "a bridge too far".

Now, however, the US National Science Foundation (NSF) has decided that it's time to span the great divide between neuroscience and education. Over the next five years, it is giving more than US\$90 million to four large, multidisciplinary teams incorporating cognitive neuroscientists, psychologists, computer scientists and educationalists. A further series of grants will be announced later this year.

Rather than making the simplistic connections that so irked Bruer, these teams want to give the craft of teaching a solid scientific underpinning. They aim to devise practical

teaching methods that will complement the brain's natural development, in part by integrating advances in cognitive neuroscience with cutting-edge information technology.

"The science of learning is ripe for a breakthrough," claims Andrew Meltzoff, co-director of the Institute for Learning and Brain Sciences at the University of Washington in Seattle, which is leading one of the four NSF-funded teams. "And the way it will occur is to build a bunch of 'mini bridges' from one discipline to another, instead of one big unsupportable bridge that goes all the way from neuroscience to education."

Treacherous territory

Maybe so, but Meltzoff and his colleagues are stepping into treacherous territory. Education is a highly politicized field, experts warn, littered with obstacles to reform and populated by powerful individuals with their own pet theories. "Even with good ideas, getting them into the classroom requires you to jump over all these political hurdles, some of which you have no control over," cautions Pamela Clute, a specialist in mathematics education at the University of California at Riverside, who spends much of her time working with

elementary and high-school teachers.

Until now, science and educational research have not mixed well. Lacking common measurement standards, educationalists have touted theories that are more like philosophies, says David Klahr, a psychologist at Carnegie Mellon University in Pittsburgh and a member of another of the NSF-funded teams. "In education it's like: 'Here's my theory. I think kids can do this or can't do that,'" he complains. "So theory is very mushy."

As a result, the field evolved into camps of specialists fighting to advance one theory of learning over another. Meanwhile, neuroscientists were holed up in their labs testing the ability of new imaging tools to deliver clues about which areas of the brain are involved in key aspects of learning. In another intellectual ghetto, computer scientists were busy using neural networks and fancy algorithms to model learning. Pity the poor teachers who were left trying to make sense of it all, barraged with brain-based pseudo-theories with no credible basis.

The challenge was to get everyone working together. So two years ago, the NSF asked researchers to come up with proposals to address fundamental aspects of learning, each

of sufficient size and scope to warrant a grant of more than US\$20 million. Each proposal also had to reach directly into the classroom or another real-world setting.

Now the first four teams to be selected are getting down to work. At Dartmouth College in Hanover, New Hampshire, neuroscientists are using magnetic resonance imaging (MRI)

to visualize the activity of children's brains as they learn. A collaboration led by the University of Washington is investigating how the brain learns in a variety of settings. At Carnegie Mellon University, meanwhile,

computer scientists and psychologists are leading a team that is refining computer-aided teaching tools. They are also setting up a 'LearnLab' in which teachers can get involved in the group's research. And a team headed by researchers at Boston University is creating textbooks and modules that focus on the mind and how it learns — with the goal of training teachers, as well as students.

The Dartmouth team has the strongest neuroscience component. Team leader Michael Gazzaniga is using a technique called diffusion tensor imaging, which analyses MRI data to determine how efficiently different parts of the brain are wired together². He is investigating

differences between children who are fast or slow readers and, in unpublished work, has already found that connections between the two halves of the brain are much more direct in children who integrate visual information quickly — a skill that might correlate with better reading.

"We are peeking under the hood at something that underlies this ability to read," Gazzaniga says. By working out the differences in brain organization that underpin children's reading abilities, he hopes it will

eventually be possible to determine which teaching methods promote the formation of the most efficient neural connections.

Also at Dartmouth, Daniel Ansari is investigating numeracy. Ansari (pictured, left) suspects that young children have two systems for doing mathematics. One system, called magnitude representation, is the crude ability to estimate quantity. Children younger than two years can usually distinguish between a large pile of candy and a small one, for instance. Ansari is testing toddlers with cards showing dots and pictures, to see if those who have trouble judging magnitude might be the ones who, later in life, will struggle with counting,

addition and, ultimately, higher mathematics.

The other system, called exact number representation, involves taking the idea of a large pile versus a small pile and coming up with numbers to stand for pile size. The ability to do this marks a huge leap in cognition, says Ansari, who is trying to visualize what this development looks like in the brain. Very young children won't lie still in MRI scanners, and so can only be studied using behavioural tests. But Ansari is running MRI scans on children aged seven to twelve, and hopes to distinguish how and when a child's exact number sense begins and whether early failures to develop the skill mark those who will subsequently struggle with mathematics.

Once Ansari has figured out the brain regions involved in learning to count and estimate numbers, he can start devising exercises to facilitate the process that could one day be adopted in preschool. Older children who are having trouble with mathematics, meanwhile, might be given tests to see whether their problems lie with their sense of exact numbers, and potentially be set similar exercises to bring this basic skill up to scratch.

Adolescent emotions

After children have entered adolescence, emotions become a major influence on their ability to perform at school. This is being investigated by Abigail Baird, another member of the Dartmouth team. She is using MRI to watch what happens in the brain when a student hears a nasty comment or a gender slur and then proceeds to flub an exam. Outreach to teens who are struggling with their emotions is central to this work: Baird and her undergraduate assistant Jane Viner have created a 10-week mentoring programme that helps teenage girls cope with aggressive interactions. Girls who took part subsequently showed an increase in the activation of their prefrontal cortex — an area of the brain thought to help rein in our emotions — suggesting that they had indeed acquired some new cognitive strategies.

In addition, the entire Dartmouth group is connecting its work with real-world education through a link with Steve Michlovitz, director of curriculum at the Windsor Central Supervisory Union, which oversees schools in Woodstock, Vermont. Formerly a school teacher, Michlovitz became interested in brain research while teaching a graduate course for education students at a local college, and more than a decade ago asked if he could bring his classes to Dartmouth to observe neuroscientists in action. Michlovitz now educates teachers about how the brain develops and functions, encouraging them to be more critical about reports on the latest findings and to

"We are peeking under the hood at something that underlies the ability to read."
— Michael Gazzaniga

IMAGE
UNAVAILABLE
FOR COPYRIGHT
REASONS

Number crunch: toddlers who have trouble judging quantities may go on to flunk maths at school.

INSTITUTE OF CHILD HEALTH, UCL

B. ERLANSON/GETTY

IMAGE UNAVAILABLE FOR COPYRIGHT REASONS

Winning formula: pupils who use cognitive tutor software, which adapts to individual needs, learn algebra faster than those receiving ordinary tuition.

avoid the sort of pseudoscience criticized by Bruer. Every two years he organizes an educational conference that brings together parents, teachers, administrators and cognitive scientists from Dartmouth and elsewhere.

Outreach is also central to the plans of the Pittsburgh Science of Learning Center, based at Carnegie Mellon, where artificial intelligence researchers are building on their success in developing an automated 'cognitive tutor' for algebra. After dissecting the steps to learning algebra, a team led by Ken Koedinger devised algorithms that play the role of a tutor, doing set problems along with a student and offering hints. By scoring the student's answers, the cognitive tutor recognizes when to stop giving so many hints, and eventually withdraws completely — until the student makes a mistake or asks for help.

Pupils who use the tutor learn algebra 50% faster than those taking regular classes, and also score 10–25% higher on standardized tests³. Today, the tool is being used at 1,800 schools across the United States.

With the new NSF funding, Koedinger and his colleagues are teaming up with experts in other subject areas to build six more cognitive tutors in subjects including geometry, chemistry and foreign languages. They also want to have their tutors make smarter decisions about when to step back and let students try problems on their own.

But the most ambitious part of the Pittsburgh project is the LearnLab, an experiment in outreach that allows teachers and students to work with the cognitive tutors, which will tailor themselves to complement the methods employed by each teacher. Teachers are also being encouraged to get involved with learning experiments, providing a framework to

conduct a wide range of educational research projects without having to build relationships with each school from scratch. "When a researcher has a great idea and they want to go do it in a school, they currently have to start out cold and get permission," Klahr says.

Even with such outreach, implementing the findings of the NSF programme will be daunting. The hurdles include cost and entrenched curriculum standards. In California, for example, a recently passed law forces districts that adopt certain textbooks to teach from them — with no deviation — in order to receive federal dollars. "Those who do education research need to go out in the real world and really see what it is going to take to implement their ideas," says Clute.

And then there are the teachers. In order for innovation in teaching to catch on, they have to grasp the ideas involved. Clute points out that, in California, more than half of middle and high-school teachers have no undergraduate training in the subject they are teaching. So how can reformers ever expect to succeed?

One NSF centre, based at Boston University, is tackling this problem. By creating textbooks and teaching modules that focus on learning and the mind, researchers hope not only to educate students about cognitive science, but also to open teachers' eyes to the potential of findings emerging from the NSF programme. "People have a deep hunger to understand themselves better in such a complicated technologically driven world," says Stephen Grossberg, who heads the Boston centre.

The final collaboration, between the University of Washington and Stanford University in California, is taking the science of learning out of the classroom. Meltzoff plans to team up with ethnographers — who study people in

their own cultural contexts. They will visit families and hang out in playgrounds to discover how children learn outside school.

He is particularly interested in the development of rational thinking, which can be studied by looking at whether children recognize that weighing up lists of pros and cons is a better way of making a decision than flipping a coin. By the age of eight, children usually say that coin-flipping is inferior, but will often revise their opinion if told a story in which flipping a coin led to a good result. Meltzoff suspects that conversations at home have a strong influence on the development of rational thinking, and will use the new NSF funding to put this idea to the test.

Explosion of findings

While individual projects such as Meltzoff's may seem arcane in their approach, he believes that the NSF programme in its totality could transform educational research. "There is a groundswell here about trying to understand the science of learning, bringing together practice and science," he says. "We are going to see an explosion of interdisciplinary findings that we have not had before in learning science."

Bruer, once the arch-sceptic, agrees that real collaboration across the various disciplines represented in the NSF initiative could lead to advances. He even chairs the group that oversees the University of Washington team. But he still cautions against over-enthusiasm, especially given the seductive power of new brain-imaging technologies. "You never can tell where research is going to lead," he says. "But the danger to everyone, the NSF in particular, is expecting too much, too soon."

Trisha Gura is a freelance science writer based in Boston.

"Education researchers need to go out into the real world and see what it will take to implement their ideas." — Pamela Clute

1. Bruer, J. *Educ. Researcher* **26**, 4–16 (1997).
2. Baird, A. A. et al. *J. Cogn. Neurosci.* **17**, 687–693 (2005).
3. Koedinger, K. R. et al. *Int. J. Artif. Intell. Educ.* **8**, 30–43 (1997).

BUSINESS

Array of possibilities opens up in genotyping

The market for microarrays — chips designed to test biological samples for genetic content — is heating up, fuelled by their growing use as tools for genotyping, or pinpointing the genetic variations behind a predisposition to disease. And industry leader Affymetrix is manoeuvring to assert its dominance.

On 31 May the inventor of the gene-chip, based in Santa Clara, California, said it would purchase ParAllele, a San Francisco company selling genotyping kits that work with Affymetrix microarrays, for \$120 million.

Advances in preparing samples for microarray analysis are increasingly important to researchers who want not only to search whole genomes at one go, but also to target specific regions of them. ParAllele has specialized in this area since it was co-founded in 2001 by Ron Davis, head of Stanford University's Genome Technology Center, who also had a hand in Affymetrix's creation back in 1993.

ParAllele sells sample preparation kits that help researchers use microarrays to detect single nucleotide polymorphisms (SNPs), genetic variations where only one base-pair differs. Its best-known tool, the molecular inversion probe (MIP), allows for the screening of up to 20,000 genetic markers at once. Even more importantly, it enables researchers to customize panels of SNPs to target specific regions of the genome. Among other things, that will help geneticists study variations in disease susceptibility between groups of different ethnic origin.

"Technology has certainly been limiting," says David Altshuler, head of the medical and population genetics programme at the Broad Institute in Cambridge, a joint project of Harvard University and the Massachusetts Institute of Technology. "The scale of whole-genome studies is hundreds of thousands of SNPs in thousands of individuals, and that simply hasn't been technically or financially approachable until this year."

ParAllele chief scientist Tom Willis describes the takeover of the company he helped found as a "bittersweet" moment. "Every successful start-up, I imagine, has a belief in its vision that must border on the irrational," he says, adding that ParAllele never intended to compete fully by building its own microarrays. "It just made sense to partner with Affymetrix," he says.

IMAGE
UNAVAILABLE
FOR COPYRIGHT
REASONS

H. KING/CORBIS

Group dynamic: ethnicity may play a role in disease.

ParAllele's decision to sell out to its larger rival reflects a general trend for biotechnology companies, says San Diego-based industry analyst John McCamant, editor of the *Medical Technology Stock Letter*. The Santa Clara company is now "the 800-pound gorilla in the room", says McCamant, noting that competitors — such as San Diego-based Illumina — may have to reconsider their options.

But Illumina, whose technology uses fibre optics embedded in glass beads to detect SNPs, says the deal came as no surprise. "We don't think the competitive dynamic will be very different" as a result of it, says Jay Flatley, its chief executive.

Affymetrix sales in 2004 were \$345 million, and its stock held steady at \$52 on the acquisition announcement, valuing the company at about \$3.3 billion. Illumina, which was founded in 1998, had sales last year of around \$50 million. Its stock has doubled in value since its February acquisition of a rival bead array maker, CyVera Corporation of Connecticut.

Biologists, meanwhile, expect that rapid technical progress in the sector will continue to provide them with better tools. Given Willis's assurance that key scientific staff will stay on at Affymetrix, Thomas Hudson, a geneticist at McGill University in Montreal, expects their development work to continue apace.

Richard Gibbs, director of the Baylor College genome sequencing centre in Houston, Texas, predicts a burst of activity in disease-gene discovery. "Technology is not the bottleneck any more," agrees Greg Yap, an Affymetrix vice-president for marketing.

Virginia Gewin

"The scale of whole-genome studies simply hasn't been approachable until this year."

IN BRIEF

AIDS VACCINE DUO The New York-based, non-profit International AIDS Vaccine Initiative (IAVI) is to partner pharmaceutical giant GlaxoSmithKline in an effort to develop an AIDS vaccine.

The partners say they will pursue a vaccine based on an adenovirus, which is stripped of its infectious genes, focusing on those that naturally infect chimpanzees. Previous vaccine candidates have been based on adenoviral vectors that infect humans.

Neither party will disclose the cost of the collaboration, but it is expected to run to several million dollars and is seen as an important step in establishing new avenues in the search for an effective AIDS vaccine.

PATENT AVALANCHE Political uncertainty surrounding the future of embryonic stem-cell research isn't preventing a veritable rush of patenting activity, a survey of the biotechnology industry has found.

Marks & Clerk, a London-based consultancy, says that patent filings related to stem cells rose by four-fifths between 2000 and 2004, during which time more than 3,000 patents were filed on stem-cell technology worldwide.

Despite restrictions on government funding for embryonic stem-cell research, US patents outnumbered those filed in the three next most active countries — Australia, Britain and Japan — by four to one.

EASTERN TIE-UP Tokyo-based Gene Networks International said on 20 June that it will acquire Shanghai Genomics, a Chinese start-up company that is developing a drug to treat fluid build-up in the lungs.

The Japanese company said that the cross-border acquisition — the first venture of its type by a Japanese biotechnology company in China — will enable it to combine research in both countries with drug development in China, where clinical trials are faster and cheaper.

The combined operation will have about 100 employees, and hopes to raise money in the future through a public share offering in Japan.

Promoting dialogue is the best way to combat ID in classrooms

SIR — Even after more than a century of vindication for evolutionary biology, creationism remains a stubborn problem for science educators.

In your Editorial “Dealing with design” (*Nature* 434, 1053; 2005) and News Feature “Who has designs on your students’ minds?” (*Nature* 434, 1062–1065; 2005), theistic scientists were asked to discuss how they reconcile their faith with science. Atheistic scientists were asked to take the time to understand matters of faith and how they relate to theories and findings of modern science. I applaud these recommendations, but it seems many of my colleagues do not.

“The issue will not go away if we ignore the challenge posed by ID.”

— Herman L. Mays Jr

In the Correspondence letter from Jerry Coyne, co-signed by many of the most prominent figures in biology (“When science meets religion in the classroom” *Nature* 435, 275; 2005), this approach is rejected. Rather than expressing the neutrality of science on matters of religion, I think this letter epitomizes the disregard, if not outright hostility, towards religious faith that is all too common in the scientific community.

This refusal to discuss what some students perceive as the threat posed by evolution, and the idea that science classrooms are the place where religious views “crumble”, will only result in science teachers having to deal more with ‘intelligent design’ (ID) in the future.

I do not want to discuss religion in the science classroom any more than Jerry Coyne does, but the issue will not go away if we simply ignore the challenge posed by ID and the concerns — albeit misguided — of students. Increased student dialogue and sensitivity on the part of instructors are exactly what is called for, to combat the inroads made by ID into the science classroom.

Herman L. Mays Jr

Department of Biological Sciences,
Auburn University, Auburn, Alabama 36849, USA

Science and religion can strengthen each other

SIR — I was rather disappointed with the one-sided presentation of intelligent design theory in your News Feature (*Nature* 434, 1062–1065; 2005), and even more so with the Correspondence letters (*Nature* 435, 275–276; 2005). They take the naive viewpoint that religious and scientific thought must be in conflict. That is not the only, in fact not even

(historically) the most prevalent, mode of addressing these two important subjects.

Most of the founding fathers of western science had no difficulty reconciling their religious beliefs with their scientific pursuits. In fact, the latter grew out of the former. This discussion actually goes back as far as the writings of Thomas Aquinas and the early Greek philosophers. Some of their writings on the interplay between these two fields is far more ‘modern’, thoughtful and relevant than much published in the modern age.

As a Christian and a scientist, I have continually found that one method of pursuing knowledge complements the other. Any conflicts I have are resolved by the revelation of my errors of thought in one area or the other. Thus my scientific reasoning bolsters my faith while my religious reasoning augments my scientific pursuits.

If one believes in objective truth and is using a rational system of thought (and when properly applied, both scientific and religious thought meet these two criteria) there is no need to feel threatened by another’s pursuit of truth. In the worst case, the two groups can agree to disagree, but in the best case they can learn more about truth together, through respectful dialogue, than they can separately.

The relationship between science and religion needs attention at a fundamental level. Only if both sides sincerely try to understand each other can a mutually beneficial dialogue be resumed.

Philip C. Farese

Department of Physics, Princeton University,
Princeton, New Jersey 08544, USA

Thoughtful peer review is worth the time it takes

SIR — I read with surprise Steve Donovan’s comment, in Correspondence, that it took him “less than an hour” to review a paper (“Reviewers not attached to online submission” *Nature* 434, 956; 2005). I am not sure if this included typing up the referee’s report, but this seems a very short time indeed when it may have taken the authors a year to write the paper in the first place.

I am no expert in palaeontology, and maybe the paper Donovan refereed was relatively easy to assess. But I often have the impression when receiving referees’ reports on my work, both favourable and negative, that they were also compiled in “less than an hour”. I spend too long refereeing manuscripts — often a full week-end or more — and sometimes end up writing excessively long reports. But I need that much time to check the cited literature, perform calculations or try to interpret the data in a (generally) less fashionable way than the authors did.

Mistakes can be made by any referee, whether quick or slow, but in my experience

the revised version of an article usually benefits from more thoughtful exchanges.

I agree with Donovan that we should not make the refereeing process unnecessarily difficult. But, as it is the only way to prevent false results or erroneous claims being published, we should not reduce it to a trivial act.

Xavier Michalet

Department of Chemistry and Biochemistry,
University of California, Los Angeles,
Los Angeles, California 90095-1569 USA

Despite some flaws, online submission is the future

SIR — John P. Moore, in Correspondence, laments the increased burden placed on scientists by the many journals that use an electronic submission system (“Online submission makes authors do all the work” *Nature* 433, 800; 2005). It is true that the online system requires manuscripts to be prepared to the journals’ specifications. For some journals it goes beyond spacing or file formats to seemingly minor details such as the format and font of the citations in text, footnotes and tables, and even the number of pixels for the figures. It’s a lot of effort for first-time users, unless they are lucky enough to find someone experienced to help them. I can also imagine the problem it poses for scientists without access to all the technology.

Nonetheless, I feel strongly that the advantages of this system far outweigh the deficiencies. I can remember, twenty years ago, submitting copies of typed manuscripts by airmail from Taiwan to Europe or North America, which routinely cost half a day’s pay. I usually had to wait for a month or two to receive an acknowledgement. If I had heard nothing after three months, I would send a letter of inquiry and wait another month or so to hear if it had been received. On more than one occasion manuscripts were lost in the post and had to be sent again. Queries (in either direction) and revised manuscripts were no less vulnerable to such mishaps. Compared with this, receiving an immediate acknowledgement and being able to track progress and submit a revised manuscript online is pure heaven.

Having said that, I feel improvements can still be made. Journals could help by allowing the flexible use of the most widely used and least expensive file formats for submission. Letting authors use an edit-friendly format, rather than having to do all the formatting themselves, would increase their motivation to submit papers. Given the ever-rising cost of journals, this would be only reasonable.

Ying-Hen Hsieh

Department of Applied Mathematics,
National Chung Hsing University,
250 Kuo-Kuang Road, Taichung, Taiwan 402

BOOKS & ARTS

Expanding the Universe

Just how many dimensions are there?

Warped Passages: Unravelling the Universe's Hidden Dimensions

by Lisa Randall

Allen Lane: 2005. 512 pp. £25

To be published in the US in September by Ecco (HarperCollins).

Paul Davies

In 1884 a curious little book appeared called *Flatland*. Written by Edwin Abbott, an English mathematician, it was destined to become a minor classic. It describes a world of two-dimensional beings — entities of various shapes for whom 'up' and 'down' have no meaning. The book describes their baffling experiences when presented with three-dimensional processes that penetrate their universe. Apart from its entertainment value, *Flatland* serves to educate the reader in the possibility that our three-dimensional world might in fact be merely a 'slice' or section in a higher-dimensional space to which access is restricted. That is, there might be a direction in space, which we can neither see nor move through, that points at right-angles to the familiar perpendicular directions (latitude, longitude and altitude). The basic message of *Flatland* is that just because we can't visualize higher dimensions, it doesn't mean they don't exist.

Mathematicians have long known that spaces of more than three dimensions make good logical and mathematical sense. Indeed, they are routinely used as a calculational device in science and engineering. In recent years, some theories of fundamental physics have postulated that additional dimensions of space really exist in the physical world. *Warped Passages* explains the rationale behind this seemingly extravagant claim.

There are two distinct ways in which one or more dimensions might be concealed from our senses. They could be rolled up to a very small size, just as a garden hose, which looks like a wiggly line from afar, turns out on closer inspection to be a two-dimensional sheet rolled into a thin tube. The idea that what we perceive as a point in space is in reality a tiny circle around a fourth dimension dates back to attempts in the 1920s to provide a unified geometrical description of gravitation and electromagnetism based on Einstein's general theory of relativity. These days physicists are more ambitious and are trying to incorporate



Like Jackson Pollock paintings, specialized science books can provide a rewarding cultural experience.

all the forces of nature, including the weak and strong nuclear forces, in a grand unified scheme. The most promising candidate, string theory, describes all matter as little loops of string that vibrate in ten-dimensional space-time, of which six dimensions are rolled up in a complicated shape.

The other way to achieve hidden extra dimensions of space is to suppose that some sort of physical force barrier confines all normal matter, together with the light by which we see the world, to a three-dimensional 'membrane' embedded in a four-dimensional 'bulk'. The author of *Warped Passages*, Lisa Randall, is an expert in these 'brane' theories, which emerge naturally as by-products of string theory.

These outlandish ideas would be pointless speculation unless there was some way for the extra dimensions to manifest themselves. String theory predicts that at energies a billion billion times those attained by particle accelerators, observable things would happen to the strings in the rolled-up dimensions. Brane theory is less demanding; it has been conjectured that the Large Hadron Collider (LHC), currently being built at the CERN particle-physics laboratory near Geneva, might reveal

traces of a fourth space dimension. This is because gravitation is not confined to branes but leaks out into the bulk. Some versions of the theory predict that, as a result, mini black holes and wormholes could conceivably be created from the high-speed collisions between the LHC's protons and antiprotons.

Randall has to confront an uncomfortable fact that faces most authors who tackle sweeping ideas of unification in fundamental physics. The book's peg — the possible existence of extra dimensions in space — is easy enough to explain. But motivating the conjecture requires a grand tour of some of the toughest and most abstract topics in science. To make her case for higher dimensions, Randall must first explain the general theory of relativity, with its notion of warped space-time. She then needs to set out the standard model of particle physics, with its confusing proliferation of subatomic objects — quarks, leptons, bosons of various sorts — and the details of the forces that act between them. Then she has to go into esoteric topics such as spontaneous symmetry breaking, Higgs potentials and supersymmetry. Because unification physics is formulated in terms of quantum mechanics —

or rather, the quantum theory of fields, a much more demanding extension — that has to be covered too. Randall must do all this before she can properly describe string theory, and must then proceed to a more obscure generalization called M-theory. Furthermore, the exposition must be crafted without the explicit use of mathematics. Only then can Randall explain how branes emerge from a welter of concepts that are largely unfamiliar to the general reader. She achieves this tour de force by cutting quite a few corners, but even so, this is a long and dense read by any standards.

I suppose it is in the nature of unification that one can't leave anything out. After all, the

thrust to create a final 'theory of everything' sprang from a desire to bring all physical theory into a common conceptual scheme. Randall makes a heroic attempt at completeness, but I wonder how many non-physicists will stay with the narrative to the bitter end.

A paradox of 'popular' science publishing is that even highly technical and specialized books still sell well, especially when there is a new author or a new twist to the subject that attracts headlines. I suspect that these books are rarely understood. Perhaps readers don't really intend to follow them studiously, but wade through the expositions as a cultural experience, rather like reflecting on a Jackson

Pollock painting — you know it's very clever and you assume it means something profound to the creator, but it belongs to an alien world. Maybe readers of popular physics and cosmology books simply enjoy being perplexed, or are satisfied merely to glimpse, voyeuristically, the arcane world of the theoretical physicist without actually having the foggiest idea of what is going on. Whatever it is that attracts these readers, *Warped Passages* has plenty of intriguing material to keep them happy. ■ Paul Davies is at the Australian Centre for Astrobiology, Macquarie University, Sydney, New South Wales 2109, Australia. His latest book is *How to Build a Time Machine*.

Making small talk

The Language of Life: How Cells Communicate in Health and Disease by Debra Nihoff

Joseph Henry Press: 2005. 306 pp. \$27.95

Fran Balkwill

"The language of all cellular societies is similarly based not on sounds or gestures but on chemistry. Using molecules where we would use words, constructing sentences from chains of proteins, the cells that make up...multicellular organisms inform, wheedle, command, exhort, reassure, nurture, criticize and instruct each other." This is how Debra Nihoff introduces *The Language of Life*, her original and absorbing book on the language of modern biology.

At its best, the prose is quite breathtaking. The first chapter immerses us in the world of a waltzing, shuffling, twirling *Escherichia coli* bacterium as it moves towards a potential meal. The process of chemotaxis is beautifully elucidated, with wonderful analogies to explain receptor-ligand interactions. Take for instance this personal ad: "Are you my better half? — SFP (single folding protein) with secure position in healthy cell seeks compatible molecule with interest in chemical engineering, architecture, or communication for exclusive short-term relationship."

Nihoff excels when describing the chemical language of bacterial colonies, neurons, the immune system and the first eukaryotes. Indeed the 16 pages describing the innate and adaptive immune systems are an almost perfect example of how to communicate complex scientific ideas to a non-specialist audience.

Cell signalling is a complex subject, especially as it occurs in the cells of vertebrates, and the general reader may struggle in some sections of the book. However, after each 'tough' molecular-biology section, Nihoff glides easily into simple analogies, or even family anecdotes, and manages to keep a reader's interest. The descriptions of cell communication are enlivened by stories of key scientists

and their discoveries, and by interviews with some of those currently contributing to the field. This important subject is moving quickly as scientists rapidly decipher the language of cells. It has a strong translational element and drives pharmaceutical research.

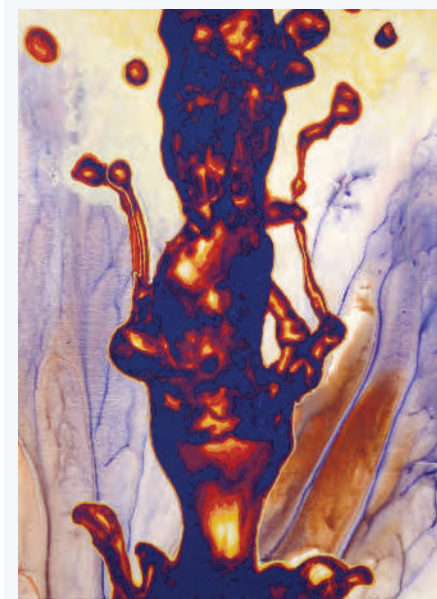
The 'disease' aspect of the title mainly relates to cancer and autoimmunity, where basic research has already led to novel, targeted and effective therapies. To cover all this ground, the book needs to be both wide-ranging and topical — and it is, impressively so. Nihoff has done her research well.

Biology is a continuum, but sometimes this point is not made strongly enough. Readers might benefit from more frequent reminders of the parallels between primitive and more sophisticated cellular systems. Although the final chapter on the 'virtual cell' is important, I would have appreciated a concluding chapter that brought the whole story together, emphasizing the common lines of communication

that link higher organisms with marine bacteria and the Hawaiian bobtail squid.

Having said that, this book is an invaluable introduction to, or reminder of, biomedical science for many lab scientists, and will also be enjoyed by a general scientific audience. Many *Nature* readers spend their working days trying to listen to the language of cells and to find ways to stop cells talking so much. Will these readers enjoy this book? My own personal benchmark for a 'popular' science book is that it is placed ahead of books by Ian Rankin, P. D. James and Patricia Cornwell as favourite bedtime reading. Nihoff's book passes that test. I shall certainly take it into the lab as recommended reading for my PhD students, and will be interested to hear their views. ■

Fran Balkwill is at the Cancer Research UK Clinical Cancer Centre, Barts and the London, Queen Mary's School of Medicine and Dentistry, John Vane Science Centre, Charterhouse Square, London EC1M 6BQ, UK.



EXHIBITION

Colour vision

"The freedom and responsibility of selecting a palette for an image is both mine and the original scientists'," says Jonathan Feldschuh, a painter based in New York who works with scientific images. His recent inspiration has come from simulations that visualize complex data, especially those representing phenomena that cannot be observed directly.

Scientists like to use colours objectively to represent information in such cases. Feldschuh also uses colour, to provide an aesthetic element, as in this image *Drop Formation #2*. Its warm tones contrast with its shapes, which are reminiscent of nuclear explosions.

'Simulations', an exhibition of Feldschuh's work, can be seen at the Cynthia Broan Gallery in New York until 9 July.

A.A.

IMAGE UNAVAILABLE FOR COPYRIGHT REASONS

Hitching a lift

Out of Eden: An Odyssey of Ecological Invasion

by Alan Burdick

Farrar, Strauss & Giroux: 2005. 336 pp.
\$25

Gabor Lövei

When geographical barriers are removed and species invade new areas, the potential consequences for biodiversity are grave. If all geographical barriers were removed, the number of mammal species would be expected to fall from 4,200 to just 2,200, for example. Unfortunately, removing geographical barriers is precisely what we are doing. Unprecedented traffic of people, goods, materials and (often unintentionally) invading organisms circulates the globe.

In a vivid illustration of this 'propagule pressure', Alan Burdick's book *Out of Eden* describes a recent voyage using traditional Polynesian sailing vessels. On their way home to Hawaii, the canoeists reported the appearance of a painfully biting insect. The entomologists they consulted concluded that some tiny biting midge (common name: 'no-see-um') must have boarded the canoes at the Marquesas Islands. None of the species it might have been was present in Hawaii, so the quarantine authorities ordered a thorough cleaning of the boats before they docked. Spraying three times with insecticides, scrubbing the boat four times, throwing everything organic overboard, and keelhauling the sails got rid of the midge, but the following stowaways still survived: "four species of fly, two species of ant, a cockroach, two spiders, a book louse, a parasitic wasp, a beetle, several snails, some live shrimps, a gecko, two species of eye gnat, and a scale insect that in some parts of the world is considered a serious agricultural pest."

Out of Eden chronicles the author's visits to

Guam to get a first-hand impression of the effect of, and research on, the "snake that ate Guam", the brown tree snake *Boiga irregularis*.

The next stop is Hawaii, which is under threat of invasion by the same snake and many other species. Several people studying various exciting phenomena appear, and it seems that the uniqueness of the Hawaiian biota moulded their personalities, too (or are we faced with pre-adaptation?). There are some fascinating stories (my favourite involves some singing underground leafhoppers), and Burdick drip-feeds a lot of complex information. Some topics, such as the niche concept and the 'empty niche' problem, are very well explained. The discussion of ecological disturbance attributed to feral pigs in Hawaii provides an interesting glimpse of the cultural conflicts surrounding invasion biology.

In the third part of the book, several detailed chapters describe the threat of marine invasions. As a terrestrial ecologist, I enjoyed reading these chapters and found them very informative; however, lay readers might think they contain too much biological detail. Nevertheless, this emerges as the best-written and most original part of the book.

This is journalist Burdick's first book, and unfortunately this sometimes shows in the style. I was slightly disturbed by the exaggerated statements, warfare-like language (even though there is a fine discussion of the consequences of using such language in invasion biology), occasionally cheap *bon mots* ("If the selection pressure is applied by an alien species, is it still 'natural' selection?") and a constant breathlessness. What may have a place in a magazine article becomes tiring over the course of a book — one cannot remain permanently excited for more than 300 pages. Fortunately, as the book progresses, Burdick's style matures.

There are also some superficialities and errors, as nearly always occurs when the writer is not an expert. For example, the mathematicians Alfred Lotka and Vito Volterra worked independently and were not "a pair"; avian malaria is not a predator (which kills and consumes the prey), but a parasite (which usually does not); and nitrogen-fixing symbionts do not "gather nitrogen from soil", they fix nitrogen from the air. A few mistakes like these are not too troublesome, but here the reader is asked just too many times to benevolently glide over them. Better editing could have eliminated many of these slips.

Strangely, and inexcusably, the book has no index, references or notes, even though it is evident from the text that the author consulted a lot of published literature. This makes it almost useless for anyone who wants to go further. There are only six illustrations, and these are of mediocre quality. No reader will be familiar with all the locations and organisms mentioned, so a liberal sprinkling of (good) illustrative material should be included in such a book.

This 'odyssey' generally takes the right direction, despite occasional meanders, and has the occasional thrilling moment. Overall, however, it fails on too many details to be convincingly, triumphantly successful. David Quammen's *Song of the Dodo* (Prentice Hall, 1996) and especially Yvonne Baskin's *A Plague of Rats and Rubber Vines* (Island Press, 2002) cover much of the same territory and, where there is overlap, these are better. Nevertheless, readers of *Out of Eden* will acquire a much-needed understanding of why we should care about biological invasions. ■

Gabor Lövei is in the Department of Integrated Pest Management, Danish Institute of Agricultural Science, Flakkebjerg Research Centre, Slagelse 4200, Denmark.

Now you see it, now you don't

Cell doctrine: modern biology and medicine see the cell as the fundamental building block of living organisms, but this concept breaks down at different perspectives and scales.

Neil D. Theise

Complexity theory, which describes emergent self-organization of complex adaptive systems, has gained a prominent position in many sciences. One powerful aspect of emergent self-organization is that scale matters. What appears to be a dynamic, ever changing organizational panoply at the scale of the interacting agents that comprise it, looks to be a single, functional entity from a higher scale. Ant colonies are a good example: from afar, the colony appears to be a solid, shifting, dark mass against the earth. But up close, one can discern individual ants and describe the colony as the emergent self-organization of these scurrying individuals. Moving in still closer, the individual ants dissolve into myriad cells.

Cells fulfill all the criteria necessary to be considered agents within a complex system: they exist in great numbers; their interactions involve homeostatic, negative feedback loops; and they respond to local environmental cues with limited stochasticity ('quenched disorder'). Like any group of interacting individuals fulfilling these criteria, they self-organize without external planning. What emerges is the structure and function of our tissues, organs and bodies.

This view is in keeping with cell doctrine — the fundamental paradigm of modern biology and medicine whereby cells are the fundamental building blocks of all living organisms. Before cell doctrine emerged, other possibilities were explored. The ancient Greeks debated whether the body's substance was an endlessly divisible fluid or a sum of ultimately indivisible subunits. But when the microscopes of Theodor Schwann and Matthias Schleiden revealed cell membranes, the debate was settled. The body's substance is not a fluid, but an indivisible box-like cell: the magnificently successful cell doctrine was born.

But a complexity analysis presses for consideration of a level of observation at a lower scale. At the nanoscale, one might suggest that cells are not discreet objects; rather, they are dynamically shifting, adaptive systems of uncountable biomolecules.

Do biomolecules fulfill the necessary criteria for agents forming complex systems? They obviously exist in sufficient

quantities to generate emergent phenomena; they interact only on the local level, without monitoring the whole system; and many homeostatic feedback loops govern these local interactions. But do their interactions display quenched disorder; that is, are they somewhere between being completely random and rigidly determined?

Analyses of individual interacting molecules and the recognition that at the



Scale up: hundreds of individual ants form a superorganism.

nanoscale, quantum effects may have a measurable impact, suggest that the answer is yes. In particular, the behaviours of increasing numbers of biomolecular 'machines' are seen to rely on brownian motion of the watery milieu in which they are suspended. Previously it was thought that binding of adenosine triphosphate (ATP) and hydrolysis releases the energy that drives these tiny machines. Now, it seems that this energy is too small to move the molecular machine mechanically, but is large enough to constrain the brownian-driven mechanics to achieve the required movement. This constrained movement is neither completely stochastic (that is, brownian), nor rigidly determined (by structure or by consumption of ATP). Examples of such phenomena include actin/myosin sliding, the activation of receptors by ligand binding, and the transcription of DNA to messenger RNA.

So, at the nanoscale, cells cease to exist, in the same way that the ant colony vanishes at the perceptual level of an ant. On one level, cells are indivisible things; on another they dissolve into a frenzied, self-organizing dance of smaller components. The substance of the body becomes self-organized fluid-borne molecules, which know nothing of such delineating concepts

as 'intracellular' and 'extracellular'. The other side of the ancient argument seems to hold: the body is a fluid continuum.

Is this merely poetic description? I suggest not. The fragility of the cell as the fundamental unit has been described before as 'cellular uncertainty', akin to the Heisenberg uncertainty principle: any attempt to examine a cell, inevitably disrupts its microenvironment, thereby changing the state of the cell. But are cells fundamentally 'uncertain' or is it possible to conceive of a technology — a perfect MRI machine, if you will — that could collect the data to describe a cell completely without altering it?

Complexity analysis suggests that no machine could ever achieve this. The cell as a definable unit exists only on a particular level of scale. Higher up, the cell has no observational validity. Lower down, the cell as an entity vanishes, having no independent existence. The cell as a thing depends on perspective and scale: "now you see it, now you don't," as a magician might say.

This analysis also allows for hypothesis-based investigations of phenomena considered outside the bounds of 'traditional' biology. A prime example is acupuncture, wherein application of stimuli to special points (meridians) on the body accomplishes remote physiological effects. The meridians do not correspond to identifiable anatomical subunits. So acupuncture, although testable and useful, cannot be explained by cell doctrine and conventional anatomy.

The validity of cell doctrine depends on the scale at which the body is observed. To limit ourselves to the perspective of this model may mean that explications of some bodily phenomena remain outside the capacity of modern biology. It is perhaps time to dethrone the doctrine of the cell, to allow alternative models of the body for study and exploitation in this new, post-modern era of biological investigation. ■ Neil D. Theise is at the Division of Digestive Diseases, Beth Israel Medical Center, First Avenue at 16th Street, New York New York 10003, USA.

FURTHER READING

Theise N. D. & d'Inverno, M. *Blood Cells Mol. Dis.* **32**, 17–20 (2004).
Theise N. D. & Krause D. S. *Leukemia* **16**, 542–548 (2002).
Kurakin A. *Dev. Genes Evol.* **215**, 46–52 (2005).

D. SCOTT/CORBIS

CONCEPTS

NEWS & VIEWS

EVOLUTIONARY BIOLOGY

Males from Mars

David Queller

In an ant species — or is it two species? — females are produced only by females and males only by males. Explanations of this revelation have to invoke some decidedly offbeat patterns of natural selection.

That men and women sometimes seem like different species is the stock in trade of pop psychologists and relationship gurus. Some go even farther: men are from Mars and women are from Venus. But in reality, human sexual differences are rather small. Even a naturalist freshly arrived from Mars or Venus would have little trouble binning specimens of men with women, and not with female chimpanzees or gorillas. There are species where males and females are different enough to have fooled real earthly naturalists. But no population geneticist would be misled — males and females mix their genes in their progeny, and as a result male and female genes comprise a common, well-mixed pool. A fascinating exception to this rule is described by Fournier *et al.* (page 1230 of this issue)¹. Males and females each reproduce clonally and, like independent species, follow separate evolutionary branches.

The surprise comes from the little fire ant, *Wasmannia auropunctata*, which is hardly obscure. An invasive pest in tropical habitats, it earns a place on a list of the 100 worst alien species². Its ancestor, like other haplodiploid social insects, must have already had two other varieties of asexuality, which together set the stage for this story. Haplodiploid species produce males asexually from unfertilized eggs (Fig. 1a), so the males are haploid — they have only one copy of each gene. Fertilized eggs become diploid females with two gene copies (Fig. 1b). In social haplodiploids, environmental differences usually induce females to differentiate into one of two castes (Fig. 2, overleaf) — ‘gynes’ become reproducing queens (Fig. 1c), whereas workers (Fig. 1d) are not just asexual but non-reproductive; they pass on their genes only by helping to rear relatives in their colony, a process known as kin selection.

In *W. auropunctata*, genetic markers show that the new gynes produced by a colony are identical to the reproducing queens (Fig. 1e). The workers, however, continue to be produced sexually. This strategy, previously reported for another ant, appears to allow the queens to pass on more copies of their genes while retaining the benefits of genetic diversity in their worker force³. The strategy cuts out the

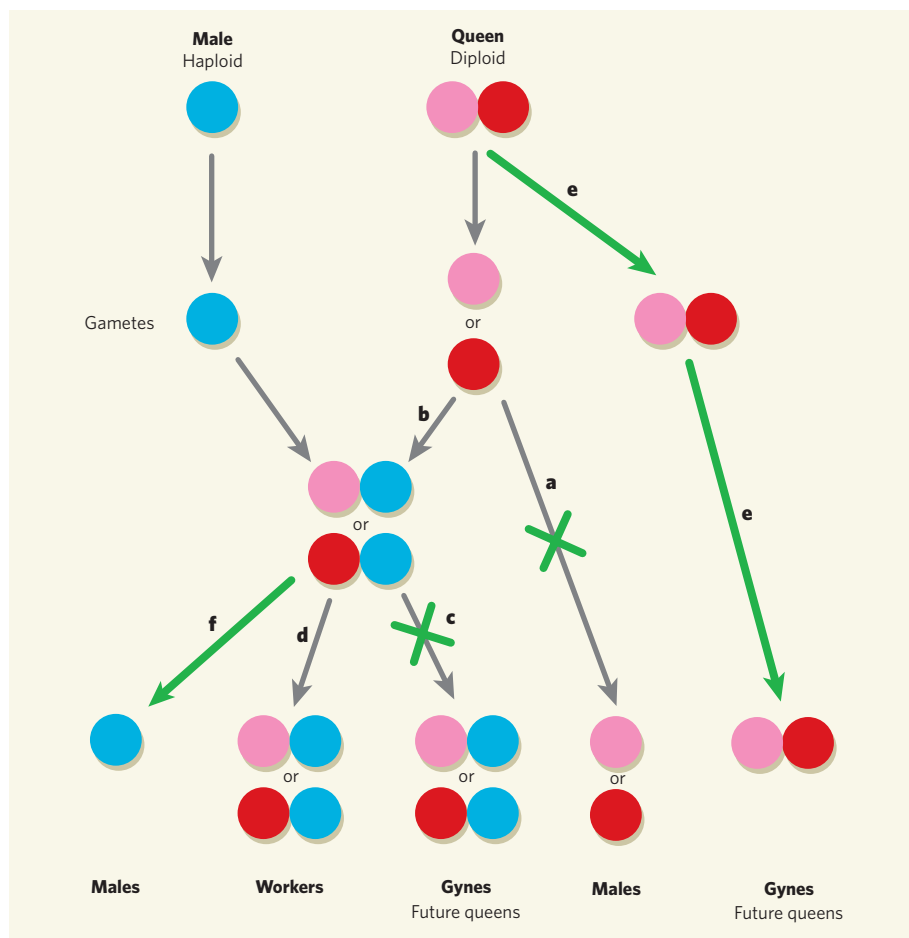


Figure 1 | Reproduction in *Wasmannia auropunctata*. Normal ants reproduce through the pathways shown by grey arrows. Each circle represents a gene (male, blue; female, red or pink). **a**, Males are produced from unfertilized eggs and therefore have one copy of each gene. **b**, Females come from fertilized eggs that can develop into either **(c)** reproductive gynes or **(d)** sterile workers. Fournier *et al.*¹ conclude that *W. auropunctata* has subtracted two pathways (green X's) and added two clonal pathways (green arrows). **e**, Gynes are genetically identical to queens, whereas **(f)** males effectively come from other males, probably by eliminating the maternal genome. The only old pathway remaining is the sexual production of sterile workers.

males as evolutionary actors — they would sire only sterile workers — but in *W. auropunctata*, the males have struck back by clonally producing other males. Fournier *et al.*¹ genotyped the sperm in the sperm storage organs of queens and found that the sperm genotype exactly matches the genotype of the males produced. They suggest that this probably results from

the paternal genome eliminating the maternal genome after fertilization, converting the diploid offspring to a haploid that will develop as a male (Fig. 1f). In *W. auropunctata*, the males seem weird enough to be from Mars.

The result is separate female and male lines that, although they combine genes in workers, never exchange them. Consistent with this, the



Figure 2 | Female castes. Two queens of *Wasmannia auropunctata*, along with workers attending to a brood. Queens and males are produced clonally, the sterile workers by sexual reproduction.

two sexes sort out into distinct branches of a genetic tree. They remain affiliated not by gene exchange but by the odd parasitic exploitation of females by males and by the mutualistic production of workers.

It is interesting to consider how this bizarre separation may rewrite the rules of colony evolution. Normally, workers evolve through kin selection based on their expected relatedness to the gynes and males they raise^{4,5}. In *W. auropunctata*, a worker's maternally derived genes — call them matrigenes (red or pink in Fig. 1) — are related to gynes but not to males. Conversely, their paternally derived patrigenes (blue in Fig. 1) are related to males but not gynes. Normally, such asymmetries are averaged out by kin selection, because worker genes can't tell when they are matrigenes and when they are patrigenes (with the possible exception of so-called imprinted genes, which have been labelled by parents, for example by tagging them with methyl groups^{6,7}).

But in *W. auropunctata*, matrigenes always remain matrigenes, so they will be consistently selected to favour gynes at the expense of unrelated males. Similarly, patrigenes remain patrigenes, and selection on them should favour related males over unrelated gynes. The result could be fierce conflict within worker genomes. Such conflict might be reduced if gynes can mate only with males of their own colony, because harming a colony's males would also harm its gynes, and vice versa.

Strange patterns of natural selection might also explain why two standard modes of reproduction have shut down. First, why would queens give up producing males by the normal pathway (Fig. 1a)? As the system stands, there is no selection for queens to produce males. A queen who produced males gains no

advantage in her own (female) gene pool; it is like putting her genes in another species. I suspect that similar logic applied, although with less force, when the female and male pools were only partially separated. The lower value of putting genes in the male pool would select against females doing so.

FLUID DYNAMICS

Impact on Everest

David Quéré

When a drop of liquid plummets onto a surface, the result is a splash — but not it seems if the process occurs at reduced atmospheric pressure. Here, perhaps, is a way to tune splash behaviour for practical ends.

A drop of liquid surrounded by air — is there anything left to discover in such a simple system, 200 years after Thomas Young and Pierre-Simon de Laplace laid the scientific foundations of capillary action? Writing in *Physical Review Letters*, Xu, Zhang and Nagel¹ reveal that air, which has been viewed as a passive fluid in the story, plays an unexpectedly active role in creating the splash that occurs when the drop hits a solid surface.

The first drops of heavy rain hitting a pond or a puddle encapsulate the full complexity of liquid–liquid impact. Hemispherical ‘storm bubbles’ (‘frozen’ by the presence of surfactants that are always present in such open environments) are distinctive features produced by the ejection of thin sheets of liquid after the shock of impact. Pioneers of high-speed imaging, such as Harold Edgerton, confirmed

Why sexually produced females cease becoming gynes may be understood by a combination of the last two arguments. First, we need to consider the patrigenes and matrigenes separately. Worker matrigenes would not gain from replacing gynes who have a copy of the gene. The patrigenes would seemingly gain, but only in the female pool. Again, it is like putting copies of its genes in another species.

If further work confirms that males and females never exchange genes and follow completely different evolutionary branches, then perhaps we really should classify them as separate species. If the females remain *Wasmannia auropunctata*, we would need a new name for the first all-male species weird enough to be from another planet. *Wasmannia mars* would serve nicely. ■

David Queller is in the Department of Ecology and Evolutionary Biology, Rice University, PO Box 1892, Houston, Texas 77251-1892, USA. e-mail: queller@rice.edu

1. Fournier, D. et al. *Nature* **435**, 1230–1234 (2005).
2. Lowe, S., Browne, M., Boudjelas, S. & De Poorter, M. *100 of the World's Worst Invasive Alien Species: A Selection from the Global Invasive Species Database* (Invasive Species Specialist Group/IUCN, 2004); www.issg.org/booklet.pdf
3. Pearcy, M., Aron, S., Doums, C. & Keller, L. *Science* **306**, 1780–1783 (2004).
4. Bourke, A. F. G. & Franks, N. R. *Social Evolution in Ants* (Princeton Univ. Press, 1995).
5. Queller, D. C. & Strassmann, J. E. *Bioscience* **48**, 165–175 (1998).
6. Haig, D. J. *Theor. Biol.* **156**, 401–403 (1992).
7. Queller, D. C. *BMC Evol. Biol.* **3**, 15 (2003).

that for any liquid or solid substrate, a corona forms upon a fast impact, which explodes into droplets to form what we usually call a splash^{2,3}.

Xu and colleagues¹ similarly use a high-speed camera (47,000 frames per second), in this case to view the impact of an ethanol drop hitting a dry glass surface at a velocity of about 3 m s^{-1} (Fig. 1). In ambient conditions, a thin sheet of liquid emerges around the impacting drop, and after 0.2 ms a corona forms, which precedes the usual splash. However, the behaviour is dramatically different in a reduced atmosphere: below 0.3 atm (a pressure still large enough to prevent the drop from evaporating), the thin film ejected after impact remains stuck to the glass. Neither corona nor splash occurs. Instead, the drop gently spreads — as if it had been placed, rather than dropped

— on the glass, except the spreading is much quicker owing to the energy of the falling drop.

This is a spectacular finding because it unambiguously demonstrates the effect of air in the splashing story. And it might also have practical consequences. The pressure reduction required to suppress the splash is modest — the air is only as rarefied as that on top of Everest. Such a partial vacuum is easily generated, providing a simple and efficient way of suppressing the emission of droplets that limits the resolution in ink-jet printing, for example. Moreover, an understanding of the role of air might also help to enhance the generation of drops in situations where it is desired, as in combustion or the production of sprays.

Xu and colleagues¹ have a surprising proposition to explain their results. As a liquid drop approaches a solid surface, the liquid can either displace or compress the air. Based on the rapid initial spreading of the liquid on the surface (the radius of the spreading disk increases as the square root of time), the authors suggest that the liquid front compresses the air. At a standard pressure, the gas resists compression, forcing the liquid edge to lift, after which this free liquid sheet cannot avoid breaking. The corresponding force (per unit area) is proportional to the gas density, speed of sound (resulting from air compression) and liquid velocity. Surface tension, which maintains the liquid cohesion, opposes this destabilizing force. Splashing occurs when the two forces are equal, and so is reduced as the gas density vanishes.

This subtle argument needs to be confirmed, because it assumes a speed on the order of the speed of sound, which is only true in the first fraction of a microsecond after impact (a much shorter timescale than observed experimentally). But it also allows us to understand why different gases behave differently; Xu *et al.* observed that splashing occurs more easily with a heavier gas (such as krypton or sulphur hexafluoride), showing an unexpected way to tune a splash.

The idea that air can be used to control the behaviour of a drop can be exploited for other purposes. A drop deposited on a pool of the same liquid will coalesce with that pool, but only after the film of air between the drop and the surface has disappeared. This typically takes a tenth of a second — a time hardly appreciable to the human eye. But if one finds a trick to renew this film, coalescence can be inhibited. Couder and colleagues⁴ have applied such a trick. Vibrating the pool at a sufficiently high frequency prevents coalescence 'for ever' (more than three days). The drop simply bounces on the surface of the pool, provided that the acceleration of the vibrating bath is at least the same as the acceleration due to gravity (9.8 m s^{-2}). The impact of the liquid on the surface is much softer than in Xu and colleagues' experiment. This maintains the integrity of the floating drop, whose mobility is enhanced through the lubricating

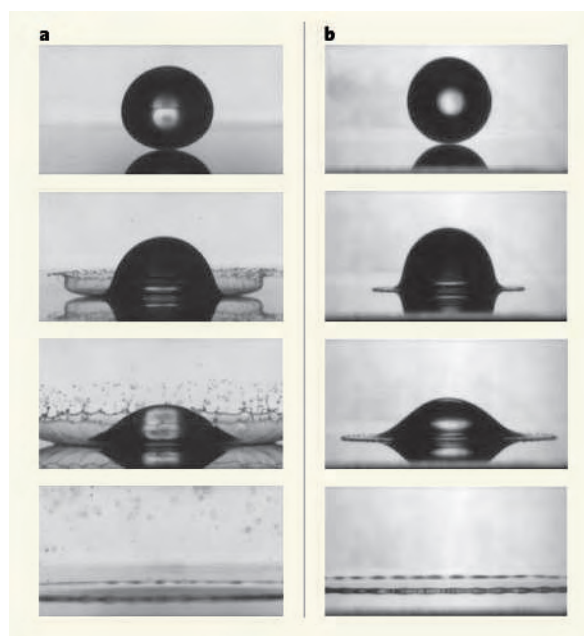


Figure 1 | Impact zone. As shown in these high-speed photographs taken by Xu *et al.*¹, air pressure determines the outcome of the impact of an ethanol drop on dry glass. **a**, At ambient pressure there's a splash. **b**, In a rarefied atmosphere — 0.3 atm, equivalent to that found at the summit of Everest — a corona fails to form and there is no subsequent splash.

effect of the air, making it easy to manipulate. By 'feeding' the drop, Couder *et al.* can maintain floating globules of liquid several centimetres across, which persist long after the vibration has stopped, because of the slow 'drainage' of air at these large scales.

The two experiments^{1,4} demonstrate that air is not as passive as we thought, and that it can be used for tuning the behaviour of drops. Our understanding of such problems as air entrapment or jet impact^{5,6} should benefit from this amended vision.

David Quéré is at the Laboratoire de Physique de la Matière Condensée, Collège de France, 75005 Paris, France.
e-mail: david.queré@college-de-france.fr

1. Xu, L., Zhang, W. W. & Nagel, S. R. *Phys. Rev. Lett.* **94**, 184505 (2005).
2. Thoroddsen, S. T. *J. Fluid Mech.* **451**, 371–381 (2002).
3. Thoroddsen, S. T. & Takehara, K. *Phys. Fluids* **12**, 1265–1268 (2000).
4. Couder, Y. *et al.* *Phys. Rev. Lett.* **94**, 177801 (2005).
5. Zhu, Y., Oguz, H. N. & Prosperetti, A. *J. Fluid Mech.* **404**, 151–177 (2000).
6. Eggers, J. *Phys. Rev. Lett.* **86**, 4290–4293 (2001).

CANCER BIOLOGY

Summing up cancer stem cells

Brian J. P. Huntly and D. Gary Gilliland

Are current cancer drugs targeted at the wrong kinds of cells? A pioneering approach to the development of treatments uses a mathematical model to follow how different types of tumour cells respond to therapy.

In this issue, Michor and colleagues (page 1267)¹ address the vexing problem that although many cancer drugs dramatically reduce the size of tumours, most cancers will eventually recur, often fatally. The authors model the dynamic changes in populations of cancer cells during treatment, and their data fit with the theory that there is a small population of cells — 'cancer stem cells' — that are ultimately responsible for the growth of tumours and are resistant to current therapies. Furthermore, the results imply that these cells may provide a reservoir for the generation and propagation of mutant cells that are also resistant to therapy. This innovative modelling strategy should allow prospective evaluation of drugs aimed at eradicating cancer stem cells.

The cancer-stem-cell hypothesis posits a very rare population of cells within tumours that are the only tumour cells with the capacity for limitless self-renewal (reviewed in ref. 2). This concept has important therapeutic implications, and may explain why it is possible to treat many cancers until the tumour cannot be detected, and yet the cancer returns. It is plausible that current treatments do not eliminate cancer stem cells, which can then regenerate the tumour once the treatment stops. This theory is not new, but it is only recently that advances in technology have allowed the prospective identification and purification of cancer stem cells, a goal first accomplished in leukaemias³, and now extended to breast cancer⁴ and tumours of the central nervous system⁵.

Michor *et al.*¹ took advantage of such advances in the understanding of chronic myelogenous leukaemia (CML) to model how leukaemic stem cells respond to therapy. CML is a blood cancer caused by a mutation that splices together the genes encoding the BCR and ABL proteins; the resulting BCR–ABL fusion protein is expressed in the bone marrow of affected individuals. ABL is an enzyme that regulates normal cell growth, but when fused with BCR it becomes permanently active, causing dysregulated overproduction of white blood cells — leukaemia.

A striking advance in the treatment of CML has been the development of imatinib, a drug that inhibits BCR–ABL activity and induces complete remission in the majority of patients with CML^{6,7}. The response to imatinib can be monitored in the clinic by measuring the level of BCR–ABL gene transcript during treatment⁸. However, several problems remain. First, as with many cancer therapies, imatinib does not cure CML. Second, some tumours develop resistance to imatinib, in most cases because new mutations result in an ABL enzyme to which imatinib cannot bind (reviewed in ref. 9).

To understand these problems better, Michor *et al.*¹ developed a mathematical model of CML based on the normal biology of blood-cell development (Fig. 1a). The model includes a hierarchy of cells at four developmental levels: stem cells, the most immature cells that can continuously renew themselves; progenitor cells, which have begun to develop

but can no longer self-renew; differentiated cells, which have begun to form a particular cell type; and terminally differentiated cells, the fully mature, specialized blood cells.

The authors then tested the model to see how well it reflected the response of CML patients to imatinib, as monitored by serial measurements of BCR–ABL transcript levels. At first, there was a rapid decline in BCR–ABL transcript number (around 5% per day) that corresponded to the model's prediction of a reduction in the differentiated leukaemic cells (Fig. 1b). Thereafter, there was a slower rate of decline of transcript number (around 0.8% per day), consistent with the modelled decrease in the leukaemia progenitor cells (Fig. 1c).

These fascinating observations are the first to characterize the kinetics of reduction of distinct subpopulations of tumour cells in response to therapy, and provide important insights into disease responsiveness and resistance. For example, the rate of increase of BCR–ABL transcript copy number after stopping therapy (Fig. 1c–e), or development of resistance to imatinib through acquired point mutations in the ABL gene, is consistent with the presence of leukaemic stem cells that have not been eradicated with treatment and are fully capable of reconstituting fulminant leukaemia. Although more work is needed to validate the model, it should be possible to test this hypothesis directly by prospective purification and measurement of distinct BCR–ABL-positive blood-cell populations during treatment.

Although there have been remarkable advances in the development of molecularly targeted drugs against cancer, such as imatinib, it is clear that additional agents will be needed to eradicate cancer stem cells — literally the root of the problem. The kinetic snapshots of how different tumour-cell subpopulations respond to treatment, derived from Michor and colleagues' model, may prove an invaluable tool as our attention shifts to the cancer stem cell as a therapeutic target. Using this model, prospective *in vivo* monitoring of response to new drugs for CML should help to identify those that strike at the leukaemic stem cell itself.

Brian J. P. Huntly is in the Department of Haematology, University of Cambridge, Cambridge CB2 2XY, UK, and D. Gary Gilliland is at Brigham and Women's Hospital and the Dana-Farber Cancer Institute, Harvard Medical School, Boston, Massachusetts 02115, USA.
e-mails: bjph2@cam.ac.uk;
ggilliland@rics.bwh.harvard.edu

1. Michor, F. *et al.* *Nature* **435**, 1267–1270 (2005).
2. Huntly, B. J. & Gilliland, D. G. *Nature Rev. Cancer* **5**, 311–321 (2005).
3. Bonnet, D. & Dick, J. E. *Nature Med.* **3**, 730–737 (1997).
4. Al-Hajj, M., Wicha, M. S., Benito-Hernandez, A., Morrison, S. J. & Clarke, M. F. *Proc. Natl Acad. Sci. USA* **100**, 3983–3988 (2003).
5. Singh, S. K. *et al.* *Nature* **432**, 396–401 (2004).
6. Sawyers, C. L. *et al.* *Blood* **99**, 3530–3539 (2002).
7. Talpaz, M. *et al.* *Blood* **99**, 1928–1937 (2002).
8. Hughes, T. & Branford, S. *Semin. Hematol.* **40**, 62–68 (2003).
9. Gorre, M. E. & Sawyers, C. L. *Curr. Opin. Hematol.* **9**, 303–307 (2002).

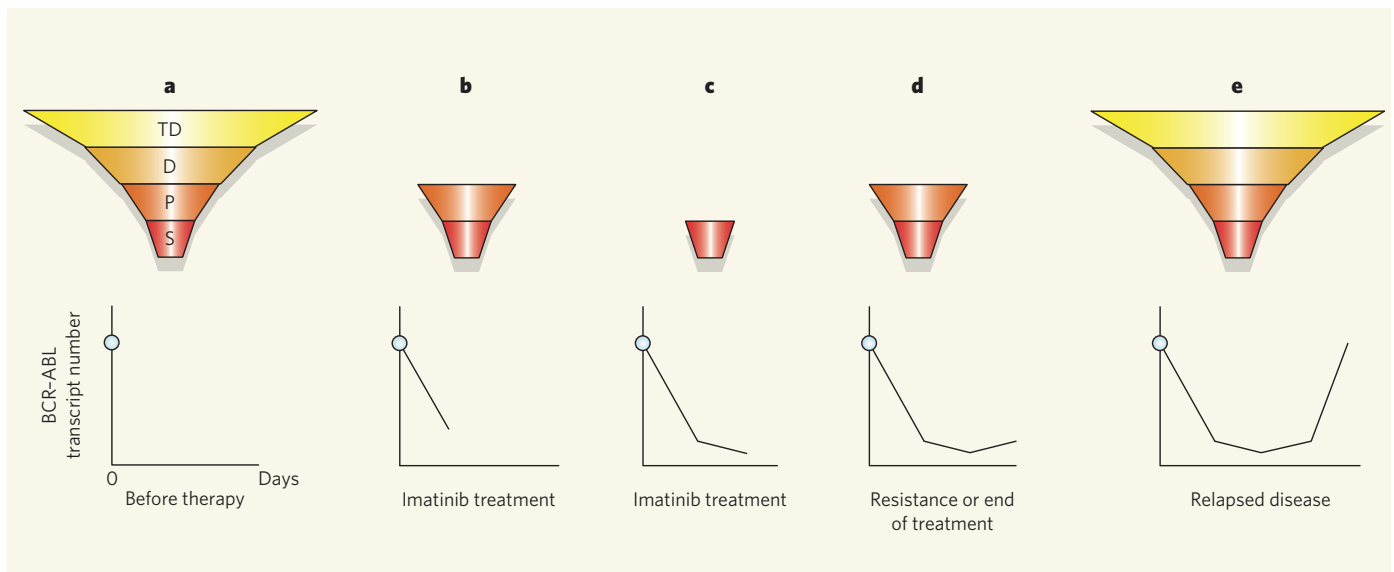


Figure 1 | Modelling leukaemia. **a**, Michor *et al.*¹ have developed a mathematical model of chronic myeloid leukaemia (CML) as four 'compartments' based on stages in the normal development of blood cells. These compartments are: leukaemic stem cells (S), which are the most immature cells and can continuously renew themselves; leukaemic progenitor cells (P), which have developed to the extent that they can no longer self-renew; differentiated leukaemic cells (D), which have begun to develop into specialized cell types; and terminally differentiated cells (TD), fully mature, specialized cells. **b**, Treatment of CML with imatinib, which inhibits the BCR–ABL enzyme that causes CML, results in an initial rapid reduction of BCR–ABL gene-transcript copy number, corresponding in the model to the depletion of the more-differentiated cells (TD and D). **c**, With continued therapy, there is a slower decline in BCR–ABL transcripts that corresponds to a reduction in the leukaemic progenitor-cell population, but the leukaemic stem cell compartment is preserved. **d**, Stopping therapy, or the development of resistance in the leukaemic stem-cell compartment, allows an initial slow rate of rise of BCR–ABL transcripts corresponding to an increase in the leukaemic progenitors, fuelled by the leukaemic stem cells. **e**, There is a rapid rate of rise in BCR–ABL transcripts that corresponds to relapse of disease. Monitoring of the rate of decline of progenitor-cell pools may aid the development of agents to target the progenitor and leukaemic stem-cell compartments in CML and other cancers.

GEOPHYSICS

Hot fluids and cold crusts

Simon Kelley

Conventional wisdom says that changes to crustal rocks pushed down deep when continents collide develop over millions of years. But it seems that some metamorphism may be caused by tectonic events lasting only a decade.

The Bergen Arcs in Norway are famous for rare and rather beautiful rocks known as eclogites. Striking, coarse-grained, and characterized by large pink garnets and a green matrix rich in silicates known as pyroxene (Fig. 1), eclogites form at extremely high pressures, and are important indicators of the conditions in the deepest parts of mountain chains. In western Norway, they formed in a continental collision some 425 million years ago; since then, glaciers have ground and polished the rock surfaces to reveal the heart of the ancient collision zone, creating a wonderful natural laboratory in which to study processes that occur deep below a mountain chain — processes that must be happening today some 50 kilometres below the Himalayas.

Geologists have learnt a great deal about continental collision-zones from the Bergen Arcs^{1–5}, and from their geologically rapid development and exhumation. But a new study challenges current understanding: based on high-spatial-resolution analyses of measurements of argon isotopes in the mineral phlogopite, Camacho *et al.*⁶ (page 1191 of this issue) propose that the partial transformation from a precursor rock-form, granulite, that characterizes the Norwegian eclogites, resulted from spasmodic, short-lived fluid-flow events lasting as little as 10 years. They also suggest that the crust at the collision zone was buried and exhumed sufficiently rapidly that it remained relatively cool during the whole cycle, which took less than 13 million years.

The eclogites of the Bergen Arcs are confined to shear zones — where rocks deform plastically as they move sideways against each other — which are also fluid pathways. Between the shear zones are regions of untransformed granulite often just tens of metres across. The Bergen Arc eclogites^{1–5} formed at depths of some 60 km and temperatures of around 700 °C. Such temperatures evoke an image of very hot and plastically deforming rocks, but herein lies a paradox: though deformed at high temperatures, the Bergen Arc eclogites exhibit features more commonly associated with tectonic processes at lower temperatures closer to the Earth's surface, such as the brittle fracturing of the garnets they contain⁷, and the formation of pseudotachylites⁸ (rocks formed by friction melting along fractures) within them.

What is more, isotopic dating using the rubidium–strontium (⁸⁷Rb–⁸⁷Sr) technique⁹ yields an age closer to the untransformed



Figure 1 | Overground laboratory. Close-up view of a Bergen Arc eclogite, the subject of Camacho and colleagues' study⁶, with its characteristic pink and green colouring.

granulite lenses in keeping with a known mountain-building event 930 million years ago¹⁰ — even though the temperatures required for eclogite formation 425 million years ago^{1–5} should have obliterated any earlier signal. The paradoxical combination of granulite preservation, high-temperature eclogite formation and the brittle features of the eclogites has led several authors to suggest that the Bergen Arc granulite–eclogite transformation occurred during short-lived fluid-flow events over less than a million years⁸. But the even shorter timescales proposed by Camacho *et al.*⁶ will make many geologists draw breath.

Camacho and colleagues used argon–argon (⁴⁰Ar–³⁹Ar) dating to measure the ages of phlogopite and amphibole mineral grains from the same untransformed granulite lenses that were investigated in the earlier ⁸⁷Rb–⁸⁷Sr work⁷. This technique works by creating the short-lived argon isotope ³⁹Ar through the irradiation of potassium (³⁹K) in mineral grains with neutrons. The age of the grains can then be ascertained from the ratio of neutron-induced ³⁹Ar to stable argon gas, ⁴⁰Ar, contained in them. (⁴⁰Ar forms from the decay of the radioactive potassium isotope ⁴⁰K, and its abundance indicates the elapsed time since the temperature was last high enough that argon could diffuse rapidly through the mineral, escaping at the boundaries between grains.) The particular advance of Camacho *et al.* is the use of an ultraviolet laser technique to measure profiles of ages across individual mineral grains ascertained using the argon–argon technique. The ages of between 820 and 895 million years that they find confirm the

rubidium–strontium results, and demonstrate just how little the granulite lenses were affected by the later eclogite formation.

The authors⁶ go on to estimate the temperature in the granulite lens during eclogite formation. Their conclusion — less than 400 °C — is a problem for the conventional interpretation of these rocks, given that a temperature of around 700 °C is required for the formation of the adjacent eclogites. Camacho *et al.* calculate that the total heating durations must have been around 18,000 years to explain the ⁴⁰Ar–³⁹Ar age profiles, but that individual fluid-flow events must have lasted just ten years to avoid significant heating of the granulite regions between the shear zones. This model evokes a radically different picture of the conditions during eclogite formation; but any alternative explanation would have to invoke a mechanism that explains why these phlogopites retained argon despite exceeding temperatures at which the gas would normally escape.

Camacho *et al.*⁶ give us a new and rapid process for eclogite formation in the Bergen Arcs.

Such rapid fluid events are not without precedent¹¹ and help to reconcile some of the high-temperature, yet brittle, features of these rocks. However, the very short timescales involved will make this idea controversial, as existing work on garnet¹² seems to indicate processes operating on a million-year timescale; but also, perhaps, simply because we geologists are attuned to thinking in millions of years, whereas the features we observe may be just the aggregations of many shorter events. There is still a lot to learn from eclogites — and Camacho *et al.* show us that the way forward is to focus more closely on high-resolution isotope variations in individual mineral grains and mineral interactions. ■

Simon Kelley is in the Department of Earth Sciences, Open University, Walton Hall, Milton Keynes, Buckinghamshire MK7 6AA, UK. e-mail: S.P.Kelley@open.ac.uk

1. Austrheim, H. & Griffin, W. L. *Chem. Geol.* **50**, 267–281 (1985).
2. Jamveit, B., Bucher-Nurminen, K. & Austrheim, H. *Contrib. Mineral. Petrol.* **104**, 184–193 (1990).
3. Boundy, T. M. & Fountain, D. M. *J. Metamorph. Geol.* **10**, 127–146 (1992).
4. Perchuk, A. L. *Petrology* **10**, 99–118 (2002).
5. Wain, A. L., Waters, D. J. & Austrheim, H. *J. Metamorph. Geol.* **19**, 609–625 (2001).
6. Camacho, A., Lee, J. K. W., Hensen, B. J. & Braun, J. *Nature* **435**, 1191–1196 (2005).
7. Austrheim, H., Erambert, M. & Boundy, T. M. *Earth Planet. Sci. Lett.* **139**, 223–238 (1996).
8. Austrheim, H. & Boundy, T. M. *Science* **265**, 82–83 (1994).
9. Kühn, A., Glodny, J., Iden, K. & Austrheim, H. *Lithos* **51**, 305–330 (2000).
10. Bingen, B., Davis, W. J. & Austrheim, H. *Geol. Soc. Am. Bull.* **113**, 640–649 (2001).
11. Van Haren, J. L. M., Ague, J. J. & Rye, D. M. *Geochim. Cosmochim. Acta* **60**, 3487–3504 (1996).
12. Erambert, M. & Austrheim, H. *Contrib. Mineral. Petrol.* **115**, 204–214 (1993).

ALICE WAIN

BIOPHYSICS

Fashionable cells

Chun Y. Seow

How can cells deform yet maintain optimal function? Probing the similarities in the properties of a cell's network of structural filaments, and those of soft glassy materials, may help in tackling this question.

Most of the cells that make up our body are supported by a cytoskeleton, an internal network of protein filaments. This network does not merely act as a scaffold that defines cell shape and organizes intracellular organelles, but also allows the cell to be malleable and motile, and adapt to strains imposed internally and externally. As they report in *Nature Materials*, Bursac *et al.*¹ have combined observations of the cytoskeleton in action with some principles derived from condensed-matter physics to offer a fresh perspective on cytoskeletal dynamics.

The need for a cell to deform while maintaining function is most obvious in the smooth muscle cell, which is embedded in the wall of hollow organs such as the urinary bladder, lungs (airways) and uterus. These organs undergo large changes in volume, implying substantial changes in cell length. Surprisingly, these drastic alterations in cell dimensions do not appear to affect cell function. The concept of plastic length adaptation was first developed for the smooth muscle of airways^{2,3}, to explain how muscle can generate maximal force over a large range of lengths. It then soon became

evident that the network of contractile and cytoskeletal filaments in smooth muscle is in a constant state of rearrangement, driven by the strains applied to the network.

Like the cytoskeleton, soft glassy materials such as pastes, colloids and foams can accommodate drastic changes of shape. Such changes occur when a cold glass, a rigid solid, is heated to a temperature at which it starts to behave as a malleable fluid. As they go about their daily business of adhering and spreading, crawling and invading, or contracting and relaxing, the cells of our body seem to orchestrate their mechanical properties in much the same way as happens in soft glasses around the glass transition temperature (which marks the crossover point between solid and fluid characteristics). But instead of changing temperature, the cell modulates something else, although with much the same effect — an 'effective temperature'⁴.

Bursac *et al.*¹ reveal further details of the behaviour of the cytoskeletal and contractile filament network of smooth muscle cells, and its striking resemblance to how soft glassy materials respond to stress and strain. Their

experiments involved attaching a microbead to the cytoskeleton of cultured smooth muscle cells from human airways, then following its spontaneous movements. They find that the microbead's movement reflects motion associated with molecular-scale rearrangements of the filament network. Most of the time nothing of great interest happens; motions are sub-diffusive, suggesting that proteins are trapped in a cage formed by weakly interacting structural proteins in a 'crowded', out-of-equilibrium microenvironment. But these sub-diffusive motions are punctuated by intermittent events thought to reflect 'hops' of the structural proteins out of one cage and into another, driven by the tendency of the discrete constituents of the crowded environment to settle slowly into a slightly more stable configuration.

An analogous hopping phenomenon has been observed in colloidal glasses⁵, and the slow settling process is known as ageing. Ageing can be reversed by a process called rejuvenation, when a soft glass is subjected to a large shear that breaks up constraints and restores the previous state of disequilibrium. Bursac

CANCER

A changing global view

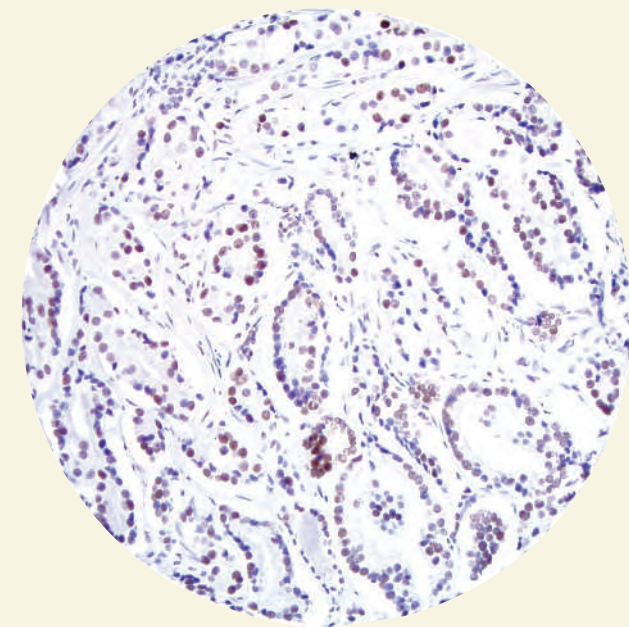
Global gene-expression profiles have emerged as a way to divide tumours that look similar into subgroups with distinct prognoses. But they can be technically demanding and difficult to implement as a routine clinical assay. Siavash Kurdistani and colleagues show elsewhere in this issue (*Nature* **435**, 1262–1266; 2005) that taking a different global view — of histone modification — can provide a similar indicator, for prostate cancer at least, that may translate easily to the clinic.

Histone proteins, around which DNA is wrapped to pack it into the nucleus, can be chemically modified in several places by the addition of acetyl or methyl groups. These modifications are reversible and can affect the expression of the associated genes. Cancer cells are known to have unusual patterns of histone modification, but so far

work has focused on individual genes and their contribution to cancer development and progression.

In contrast, Kurdistani and colleagues looked at global levels of the acetylation or methylation of five different residues in histones H3 and H4 in prostate tumour samples. They used antibodies that were specific for each modification to highlight any differences; for example, the photo here shows a section of low-grade prostate cancer tissue (Gleason score 6) stained with an antibody against dimethylated arginine 3 of histone H4.

The authors found, from two independent sets of prostate cancer samples, that histone modification patterns can forecast the risk that a low-grade tumour will recur after surgical removal. How the observed global changes in histone



DAVID SELIGSON

modifications relate to the regulation of genes relevant to prostate cancer development is not known. However, using immunochemistry to detect bulk

histone modifications in cancer samples is relatively easy, so these findings could be translated directly into prognostic markers for clinical use.

Barbara Marte

*et al.*¹ show that the cytoskeleton of smooth muscle cells can age and rejuvenate in a similar manner.

So insights gained from advances in understanding inert glassy materials can be applied to the cytoskeleton. But there are of course dissimilarities. As Bursac *et al.* point out, interactions of structural proteins depend on the energy supply from within the cell in the form of ATP hydrolysis. The energy supply and protein–protein interactions that drive the cytoskeletal reorganization are precisely regulated in cells⁶; similar active regulation does not occur in the case of glass deformation driven by strain and thermal energy.

The notion that the cytoskeleton of a smooth muscle cell is malleable^{1,4}, and that the malleability is probably precisely regulated⁷, is important for our understanding of how smooth muscle and other cell types work. Although a cell needs to be malleable to accommodate large changes in geometry, it also needs to be rigid to generate or transmit force. Some mechanisms underlying the ability of smooth muscle to plastically reorganize its contractile^{2,7} and cytoskeletal^{3,8} filaments have been proposed, but many aspects of the cell's adaptive behaviour remain unknown.

The effects of cell malleability on organ function is another area that requires investigation. It is easy to appreciate why the smooth

muscle cell needs to be malleable in hollow organs that undergo large changes in volume, but for other cell types that need is less obvious. It is likely, however, that the ability of the cell to rearrange the molecules of its internal skeleton is essential in such events as vessel narrowing, wound repair and cell migration. Abnormality in the regulation of cell malleability may therefore underlie conditions such as cell invasion in cancer, and the excessive narrowing of airways and blood vessels seen, respectively, in asthma and hypertension. The report of Bursac *et al.*¹ gives us a different way to think about how this basic rearrangement process might work. ■

Chun Y. Seow is in the Department of Pathology and Laboratory Medicine, James Hogg iCAPTURE Centre, University of British Columbia, Vancouver, British Columbia V6Z 1Y6, Canada. e-mail: cseow@mrl.ubc.ca

1. Bursac, P. *et al.* *Nature Mater.* **4**, 557–561 (2005).
2. Pratsuevich, V. R., Seow, C. Y. & Ford, L. E. *J. Gen. Physiol.* **105**, 73–94 (1995).
3. Gunst, S. J., Meiss, R. A., Wu, M. F. & Rowe, M. *Am. J. Physiol.* **268**, C1267–C1276 (1995).
4. Fabry, B. *et al.* *Phys. Rev. Lett.* **87**, 148102 (2001).
5. Weeks, E. R., Crocker, J. C., Levitt, A. C., Schofield, A. & Weitz, D. A. *Science* **287**, 627–631 (2000).
6. dos Remedios, C. G. *et al.* *Physiol. Rev.* **83**, 433–473 (2003).
7. Herrera, A. M. *et al.* *J. Cell Sci.* **118**, 2381–2392 (2005).
8. Gunst, S. J., Tang, D. D. & Opazo Saez, A. *Respir. Physiol. Neurobiol.* **16**, 151–168 (2003).

IMAGING TECHNIQUES

Particular magnetic insights

Andreas Trabesinger

Over the past 30 years magnetic resonance imaging has been refined into a widely used technique. A method known as magnetic particle imaging has now been devised which offers an inner view from a different angle.

It is fairly easy to record electrical signals from the human body, as with an electrocardiogram. The extraction of magnetic signals is a more subtle business, typically involving strong magnets and exquisitely sensitive detectors. In a magnetic resonance imaging (MRI) experiment, this effort is well rewarded, as the technique reveals fine details of a subject's inner life¹. The approach can be pushed further by introducing strongly magnetic particles into the body which selectively perturb the weak natural MRI signal and increase the image contrast². Direct imaging of these particles, however, has seemed less feasible — hence the appeal of the report by Gleich and Weizenecker³ (page 1214 of this issue), which lays out a way to do just that.

Contrast agents that incorporate magnetic particles are routinely used in clinical MRI examinations. These biocompatible 'spies', typically based on rare-earth elements or iron oxides, can highlight specific anatomical

structures, such as blood vessels or (under certain circumstances) tumours. Furthermore, they can serve as markers for processes at the molecular level⁴.

The valuable information contained in the spatial distribution of the magnetic contrast agent is recorded indirectly. In these contrast-enhanced MRI studies, the strong magnetization of the magnetic particles is used to alter the signal of the (orders of magnitude) weaker intrinsic nuclear magnetization of the body. Looking directly at the contrast agent could potentially translate its stronger magnetization into a stronger signal — or, alternatively, reduce the contrast agent required to minute amounts. However, resonance methods such as those used for MRI are often unsuitable for imaging magnetic particles, and 'inversion methods' that detect the magnetic field outside the object do not provide high spatial resolution.

Gleich and Weizenecker³ present an

approach — magnetic particle imaging (MPI) — for capturing and localizing the signal of magnetic particles inside the body. In their initial demonstration, they filled holes in a plastic plate with a commercial contrast agent (nanometre-sized particles of iron oxide, coated with dextran). When a magnetic field is applied to this assembly, the contrast agent becomes magnetized. The response to a changing magnetic field is more or less immediate, until the magnetization reaches saturation above a certain field. A further increase in field will leave the magnetization unaffected.

Now imagine that the particles in a magnetic field are further irradiated by a weak radio-frequency field (Fig. 1b–d). Saturated magnetic particles will stay in this state (Fig. 1b, d), whereas unsaturated ones will readily reply with an oscillating magnetization (Fig. 1c). As the response is nonlinear, it will contain frequencies that are different from the driving frequency of the radio-frequency field.

These signals announce the presence of magnetic particles in the body, but not their location. To add this crucial component,

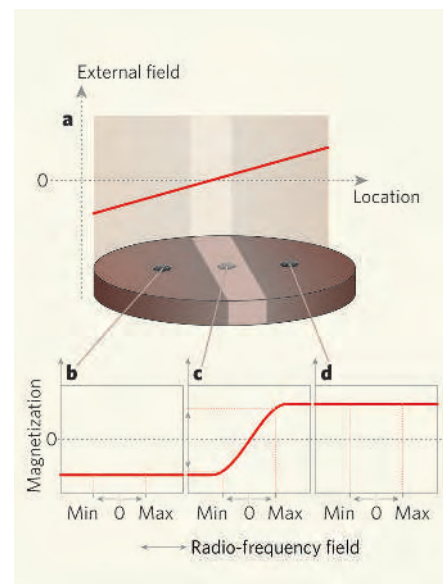


Figure 1 | Principle of the imaging technique devised by Gleich and Weizenecker³. **a**, The object to be imaged is immersed in an external field whose strength varies with location. In most regions the magnetization of a magnetic particle sitting inside the object is saturated (dark areas). **b, d**, An additional weak radio-frequency field — oscillating between a minimum and a maximum value — cannot change this state. **c**, However, in regions where the external field has a value close to zero, the additional field is able to alter the magnetization, which will start to oscillate and therefore induce a signal in a detection circuit. This signal can be unambiguously assigned to the narrow field-free region. By systematically varying the position of the field-free area in the object, a map can be created that gives the spatial distribution of the magnetic particles.



50 YEARS AGO

"The Sun", edited by Prof. Gerard P. Kuiper. — Nothing comparable with this work has appeared since the publication of Vol. 4 of the "Handbuch der Astrophysik" in 1929. A comparison of the two volumes demonstrates impressively the strides solar physics has made in a quarter of a century. The identification of 'coronium', the recognition of H⁺ absorption, postulation of the carbon-nitrogen cycle, invention of the coronagraph... the discovery of chromospheric flares and their terrestrial effects and of solar radio noise... It is regrettable that the present volume includes no contribution from the U.S.S.R.; but the cause can doubtless not be laid entirely at the door of the editor, who is as well aware as anyone how much solar research carried out in Soviet territories remains a closed book to Western readers. From *Nature* 2 July 1955.

100 YEARS AGO

It is announced in the *Times* that the Board of Trade and the Trinity House have concluded a contract with Marconi's Wireless Telegraph Company (Limited) providing for the equipment of lightships with Marconi wireless telegraph installations. This arrangement will enable the lightships to communicate with the shore and with one another by wireless telegraphy for the ordinary purposes of the lightship service, and also to report ships in distress.

ALSO

"British Bird Life". By W. Percival Westell. The wearisome procession of books on British birds drags on — a long train of volumes, all of necessity telling the same tale, and for the most part badly... At times Mr. Westell becomes ecstatic, and, blinded by the intensity of his emotions, rushes onwards regardless of obstacles — even of the rules of grammar... This book is profusely illustrated, partly by photographs, some of which are very pleasing, and partly by "original" drawings, all of which are bad.

From *Nature* 29 June 1905.

Gleich and Weizenecker allow only certain locations inside the body to send a signal. They achieve this by placing the object concerned in an inhomogeneous magnetic field which, in most regions, is strong enough to saturate a magnetic particle (Fig. 1a). Only particles situated at sites where the external field is essentially zero are not saturated, and can therefore signal in response to the radio-frequency field. By changing the location of this field-free spot (either mechanically or with auxiliary magnetic fields) the sample can be scanned bit by bit, resulting in a map of the spatial distribution of the magnetic particles.

The images obtained in the initial experiments have a resolution of well below 1 mm. This is surprising, considering that the size of the recording coils (squares with 16-mm sides) and the wavelength of the applied radio-frequency field (around 1 km) are both much larger than the size of the resolved features. MPI can be seen as a form of 'zeugmatography', a term coined by 2003 Nobel laureate Paul Lauterbur⁵ in his introduction of MRI as a concept for image formation: when two fields are combined, the first one (here, the radio-frequency field) induces an interaction with the body, and the second one (the inhomogeneous magnetic field) restricts this interaction to a limited region. In this way, there is no imposed wavelength limit and MPI can use harmless radio waves that pass through the body without significant attenuation. Furthermore, the detectors can be much larger than the smallest resolved structure, thereby opening the door to depth resolution and, ultimately, three-dimensional imaging.

Beyond this proof-of-principle demonstration, the practical usefulness of MPI remains unknown. The concept promises to complement existing methods and, in certain applications, to provide a unique internal view. MRI owes much of its versatility to the fact that our bodies (and virtually all materials) are made of nuclei that exhibit weak magnetism. MPI, on the other hand, relies on the detection of magnetic particles with stronger intrinsic magnetism, but in general those particles have to be introduced. Although MPI will reveal fewer details, it will not suffer from any background interference and should resolve structures with excellent contrast.

If the potentially higher sensitivity of MPI can be fully exploited, a fast and powerful imaging technique could be in prospect, as well as relatively cheap mobile scanners, with geometries that can be adapted to particular applications. MRI has continued to astonish us with its ever-increasing sophistication over the past three decades. MPI might offer surprises of its own.

Andreas Trabesinger is an associate editor at *Nature Physics*, based in London.
e-mail: a.trabesinger@nature.com

1. Chacko, A. K., Katzberg, R. W. & MacKay, A. *MRI Atlas of Normal Anatomy* (McGraw-Hill, New York, 1991).
2. Tóth, É. & Merbach, A. E. *The Chemistry of Contrast Agents in Medical Magnetic Resonance Imaging* (Wiley, Chichester, 2001).
3. Gleich, B. & Weizenecker, J. *Nature* **435**, 1214–1217 (2005).
4. Allen, M. J. & Meade, T. J. *Metal Ions Biol. Syst.* **42**, 1–38 (2004).
5. Lauterbur, P. C. *Nature* **242**, 190–191 (1973).

GENE REGULATION

Expression and silencing coupled

Stephen Buratowski and Danesh Moazed

The RNA interference pathway can inhibit the expression of specific genes. It now seems that an essential component of the silencing process is the gene-expression machinery itself.

Molecular biologists have been amazed in recent years by the discovery of an RNA-mediated mechanism for inhibiting the expression of specific genes — the RNA interference (RNAi) pathway. The 'RNA-induced silencing complex' (RISC) contains small interfering RNAs (siRNAs) whose sequence of nucleotide bases can pair with that of a particular messenger RNA, targeting this mRNA for destruction before it can be translated into protein¹. However, in many organisms this 'post-transcriptional' gene silencing is only part of the story: production of the mRNA can be shut off by a second siRNA complex called RITS (for 'RNA-induced transcriptional silencing').

Schramke *et al.* (page 1275 of this issue)²

and Kato *et al.* (writing in *Science*)³ now show that a gene must first be transcribed if it is subsequently to be silenced. More surprisingly, this transcription must be specifically carried out by RNA polymerase II (RNAPII), the enzyme responsible for making mRNAs in eukaryotic organisms.

DNA is packed into nuclei by being wrapped around histone proteins to form nucleosomes. RITS represses transcription by recruiting a histone methyltransferase to the target genes. This enzyme modifies histones so as to make the wrapped DNA inaccessible to the gene-expression machinery, creating a silenced nucleosome configuration known as heterochromatin. RITS requires siRNAs for its

association with chromatin, and, based on its similarity to RISC, it seemed likely that RITS would also be targeted via base-pairing of siRNAs, either to DNA or RNA¹. The dependence on transcription suggests that the target is RNA.

Schramke *et al.*² show that the transcription of the gene targeted for silencing must be carried out specifically by RNAPII. Although a polymerase from bacteriophage (a virus that infects only bacteria) can produce transcripts from a eukaryotic chromosome, silencing does not occur. Therefore, although RITS may be targeted via base-pairing to nascent RNA transcripts, an additional mechanism must exist for specifically coupling silencing to RNAPII. The two groups^{2,3} find that very different mutations in RNAPII disrupt the formation of heterochromatin: truncation of the RNAPII largest-subunit carboxy-terminal domain (the CTD, normally required for coupling mRNA synthesis to mRNA processing⁴), or a specific point mutation in the RNAPII Rpb2 subunit, both lead to loss of silencing, but with an interesting difference. Whereas siRNAs are made normally in the CTD truncation mutant, the Rpb2 mutant seems to be blocked in processing the siRNAs. Therefore, RNAPII may have multiple roles in the siRNA pathway.

Perhaps the simplest model to explain the coupling of RNA-induced silencing with transcription (Fig. 1) is that RITS is tethered to some part of the RNAPII elongation complex, which produces the target mRNA. Through this interaction, as well as recognition of specific transcript sequences, the histone methyltransferase would be localized to the appropriate target gene and so could modify the histones as the gene is being transcribed. Indeed, molecular-interaction experiments show that RITS is linked to nascent transcripts^{2,5} as well as to RNAPII (ref. 2). Furthermore, RITS and its associated factors can be chemically crosslinked to genes undergoing silencing, and this requires siRNA and the nascent transcripts^{2,5}. There is clear precedent for the coupling of transcription with chromatin modification: two other histone methyltransferases (Set1 and Set2) bind directly to elongating RNAPII and modify transcribed regions of genes⁶.

However, there are several other models that could explain why only RNAPII can mediate RITS-dependent silencing. It is possible that RITS interacts with the transcript not only through base-pairing, but also by recognizing an RNAPII-specific modification of mRNA (such as the cap structure or poly(A) tail). In this regard, it is interesting that a screen for factors that promote RNAi in the nematode worm *Caenorhabditis elegans* identified several factors that are required for proper mRNA processing⁷. The mechanisms by which these factors affect RNAi are unknown.

In addition to mRNA-processing enzymes,

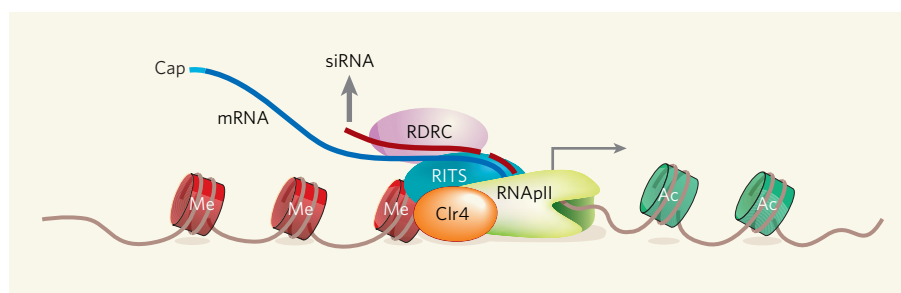


Figure 1 | Transcription-coupled RNA interference. This speculative model incorporates the results of Schramke *et al.*² and Kato *et al.*³. As RNA polymerase II synthesizes the RNA transcript, targets to be silenced are recognized by the RITS complex by siRNA-mRNA base-pairing, and possibly through direct contacts with the elongation complex. Formation of repressive heterochromatin, where histones carry a particular methylation pattern (Me) and have lower levels of acetylation (Ac), is initiated by the histone methyltransferase Clr4. At the same time, RITS and an RNA-dependent RNA polymerase complex (RDC) can generate double-stranded RNA that is processed to become new siRNAs.

the RNAPII elongation complex carries several chromatin-modifying enzymes⁶. Although bacteriophage polymerases and eukaryotic RNA polymerase III can transcribe through chromatin without disrupting nucleosomes⁸, passage of the RNAPII elongation complex leads to large changes in the chromatin⁹. Histone subunits may be exchanged as the transcription complex passes through a nucleosome. Furthermore, several transcription-dependent modifications of histones, including the methylations described above, have been identified. It may be that one or more of these alterations are prerequisites for the RITS-associated histone methyltransferase to modify its substrate.

Each of these explanations makes some testable predictions, so the process that links RITS-mediated repression specifically to RNAPII may soon become clear. However, other questions remain. For example, is transcription required only to initiate silencing, with the repressive chromatin being subsequently propagated by epigenetic mechanisms such as histone methylation? Or is some low level of transcription paradoxically required to maintain repression? Any transcripts made from a 'silenced' gene would be subject to RISC-mediated degradation, so some leakiness of transcriptional silencing may easily be tolerated. An RNA polymerase IV that apparently specializes in synthesizing siRNA precursors has recently been discovered in plants¹⁰. This suggests another mechanism for simultaneous transcription and silencing: perhaps RNAPIV can transcribe through chromatin structures that block RNAPII. It will be interesting to see whether transcription by RNAPIV is directly coupled to the silencing complex.

Another question is whether siRNAs are themselves generated during transcription, a possibility suggested by the findings of Kato *et al.*³. Once targeted by RISC/RITS, the transcript slated for destruction can be used to generate new siRNAs. This requires two enzymes — an RNA-dependent RNA polymerase that generates double-stranded RNA,

and the Dicer enzyme that cuts the RNA into siRNA lengths. Both enzymes associate with RITS or RISC complexes and could therefore be present at the site of transcription.

One of the main functions of the RNAi system is probably to suppress expression of the repetitive elements that parasitize eukaryotic genomes. A two-pronged approach to silencing makes good sense. The RISC complex can target any transcripts that manage to reach the cytoplasm from the nucleus, preventing them from being translated. But if this were the only mechanism, considerable cellular energy might still be wasted in making transcripts from the repetitive elements. The nuclear RITS complex can repress expression of those RNAs before they are even made. Because target recognition uses complementary RNA sequences, once a particular element or gene is recognized by the RNAi system, all copies in the cell will be targets for inactivation. Nucleotide sequences provide a high level of specificity, but there must be an opportunity for the target sequences to be recognized. By coupling the RNAi machinery to ongoing transcription, siRNAs can identify target transcripts as they are synthesized, resulting in efficient and almost immediate repression. ■

Stephen Buratowski is in the Department of Biological Chemistry and Molecular Pharmacology, and Danesh Moazed is in the Department of Cell Biology, Harvard Medical School, 240 Longwood Avenue, Boston, Massachusetts 02115, USA.

e-mail: steveb@hms.harvard.edu

1. Ekwall, K. *Mol. Cell* **13**, 304–305 (2004).
2. Schramke, V. *et al.* *Nature* **435**, 1275–1279 (2005).
3. Kato, H. *et al.* *Science* published online 9 June 2005 (doi:10.1126/science.1114955).
4. Maniatis, T. & Reed, R. *Nature* **416**, 499–506 (2002).
5. Motamedi, M. R. *et al.* *Cell* **119**, 789–802 (2004).
6. Hampsey, M. & Reinberg, D. *Cell* **113**, 429–432 (2003).
7. Kim, J. K. *et al.* *Science* **308**, 1164–1167 (2005).
8. Studitsky, V. M., Kassavetis, G. A., Geiduschek, E. P. & Felsenfeld, G. *Science* **278**, 1960–1963 (1997).
9. Ahmad, K. & Henikoff, S. *Cell* **111**, 281–284 (2002).
10. Vaughn, M. W. & Martienssen, R. A. *Mol. Cell* **17**, 754–756 (2005).

OBITUARY

Keiiti Aki (1930–2005)

Seismologist extraordinaire.

Everyone called him Kei, the intellectual leader who for 40 years devised new methods to achieve a more quantitative understanding of dynamic processes within the Earth. Kei Aki died on 17 May on Réunion Island, the volcanic 'hot spot' in the Indian Ocean that had been his home since retirement from academic life. He will be known for his many research results in seismology, as well as for his leadership in developing probabilistic estimates of seismic hazard.

Aki was born in Japan, and wrote that "When I was 19 years old, I applied to the Department of Geophysics of the University of Tokyo, partly because of the simplest entrance examination, only on three subjects, English, Mathematics and Physics". He moved on to seismology and was recruited by Frank Press to the faculty of the Massachusetts Institute of Technology in 1966, and thus to an academic career in the United States.

His scientific achievements include some of the earliest studies using seismic surface waves to estimate fault orientation and direction of slip at the earthquake source. In 1966 he showed how to estimate the 'seismic moment' of an earthquake from seismograms. This measure, equal to the product of fault area ruptured, average fault slip and rigidity, became recognized as the best way to characterize the size of an earthquake source and the strength of its long-period waves. Among many uses of seismic moment are integrative estimates of the earthquake-associated motion between two tectonic plates. When compared with total plate motion derived from geomagnetic, geological and geodetic methods, one learns the fraction of motion on a particular plate boundary that takes place suddenly in earthquakes — permitting estimates of the long-term seismic hazard.

In the 1960s, Aki showed that radiated seismic displacements have a 'corner frequency' below which the spectrum is flat and proportional to seismic moment, and above which it decays as (frequency)⁻². He showed that seismic coda (the waves following signal onset) can be used to make stable measurements of source spectra.

The following decade, Aki interpreted seismic-wave arrival times at a network of stations to determine the three-dimensional inhomogeneities of Earth structure beneath the network. His 'inverse method' has been applied via instrument deployments on every continent to quantify inhomogeneities in the crust and upper mantle. He also began quantitative studies of fault models

that examined the physical features controlling earthquake nucleation, spontaneous rupture, and the eventual stopping of the rupture process on a fault surface because of some barrier or asperity — or because the rupture has run out of a region of stress concentration.

As an observational science, seismology deals with vast ranges of scale. Wavelengths range over a factor of more than 10⁸; the time windows over which data are analysed range over more than 10⁹; ground displacements range over more than 10¹¹; and the size of seismic sources ranges over about 10²⁵. Consequently seismology is largely conducted as a collection of diverse specialties. Some seismologists interpret signals to map out the remaining commercial deposits in an oilfield; others quantify strong ground-shaking and measure how it decays with distance travelled. Different seismologists study the Earth's internal structure — from the inner core, up through the mantle, and on up to the shallowest structures in the crust. They identify targets for an archaeological dig, find buried pipelines, measure rates of continental deformation, monitor compliance with nuclear test ban treaties, and assess seismic hazard. Aki and his students worked in many of these specialties, and he himself retained a strong and optimistic interest in arguably the most important of them all, earthquake prediction.

In March 1975, he wrote me a letter beginning, "I wonder if you would be interested in coauthoring a text book on theoretical seismology with me..." It was a surprise, for even then he was a senior figure. I was so junior, and we were not then personally acquainted. I accepted, and will never forget our first meeting. As background to our conversation, he jotted down some apparently random descriptive phrases on a large sheet of paper that soon filled to become the working plan and table of contents for what emerged as *Quantitative Seismology*, now translated (with its more than 6,000 equations) into Russian, Chinese and Japanese. His goal was a unified survey that brought out underlying ideas relevant to many different types of seismologist.

In 1984, Aki moved to the University of Southern California, where he could experience earthquake phenomena more directly. He became the founding science director of the Southern California Earthquake Center, where earthquake geologists and the fault model of quantitative seismology were used to make probabilistic



estimates of earthquake hazards. He was also fascinated with volcano seismology, which he continued to study in 'retirement', especially on Réunion with its local seismographic network. He studied volcanic tremor, estimated the location of magma chambers, monitored magma ascent and searched for physical causes.

Aki probed more deeply into the phenomena of shaking ground, and with greater insight, than any other individual. He advanced our understanding of how earthquakes nucleate, how they reach their eventual size, and how their signals spread throughout the Earth, carrying information on their origin and what they have traversed. He showed us how to tease out this scientific information about sources and Earth structure. And he successfully faced the challenges of taking information about strong ground motions into the earthquake engineering community, and to policy-makers needing guidance on managing earthquake hazard.

Kei Aki shared ideas with his more than 50 PhD students and numerous postdocs. He helped them to shine, and many are now leaders of their field. He himself was a gentle leader, informal and approachable, pleased with the many honours that came his way but never needing them, calm in adversity, and always eager to hear of new ideas that had some support from data. With his pioneering studies of seismic moment, his results on spectral scaling and coda stability, and his demonstration of successful methods to invert data to infer Earth structure in three dimensions, he provided methods that now guide the work of thousands of Earth scientists around the world.

Paul G. Richards

Paul G. Richards is at the Lamont-Doherty Earth Observatory, Columbia University, 61 Route 9W, Palisades, New York 10964, USA.
e-mail: richards@ldeo.columbia.edu

USC/SOUTHERN CALIFORNIA EARTHQUAKE CENTER

BRIEF COMMUNICATIONS

Continuous activity in cetaceans after birth

The exceptional wakefulness of newborn whales and dolphins has no ill-effect on their development.

All mammals previously studied take maximal rest or sleep after birth, with the amount gradually decreasing as they grow to adulthood^{1,2}, and adult fruitflies and rats die if they are forcibly deprived of sleep^{3,4}. It has therefore been assumed that sleep is necessary for development and serves a vital function in adults. But we show here that, unlike terrestrial mammals, killer-whale and bottlenose-dolphin neonates and their mothers show little or no typical sleep behaviour for the first postpartum month, avoiding obstacles and remaining mobile for 24 hours a day. We find that neonates and their mothers gradually increase the amount of time they spend resting to normal adult levels over a period of several months, but never exceed these levels. Our findings indicate either that sleep behaviour may not have the developmental and life-sustaining functions attributed to it, or that alternative mechanisms may have evolved in cetaceans.

Two adult female killer whales (*Orcinus orca*) and their calves were observed for 5 months after birth (for methods, see supplementary information). Adult orcas normally show distinct periods of immobility, which occupy 5–8 hours a day under baseline conditions, while floating at the surface or lying on the bottom of pools. Figure 1a shows the suppression of typical rest behaviour in the mother for several months after birth and even less rest behaviour in the newborn over this period. The calves always rested less than their mothers and less than other adults (Fig. 1a), unlike all other mammals that have been studied. (For video, see supplementary information.)

We assayed blood concentrations of the hormone cortisol in three killer-whale mothers at 3–5 weeks postpartum, and found no significant increase above the values recorded in these same animals when pregnant (pregnant: $n = 3$, 2.6 ± 0.6 ng ml⁻¹, range 0.9–7.4 ng ml⁻¹; postpartum: $n = 3$, 3.5 ± 0.4 ng ml⁻¹, range 2.8–4.1 ng ml⁻¹). This indicates that stress was not responsible for the postpartum decrease in sleep behaviour.

Four dolphins (*Tursiops truncatus*) and their calves were also observed after birth and showed a similar pattern of sleep behaviour (Fig. 1b). Neither mothers nor calves rested at the surface during the first postpartum month, but time spent resting then gradually increased towards that spent by normal adults. As in the case of the killer whales, the neonates always showed less rest behaviour than the mothers.

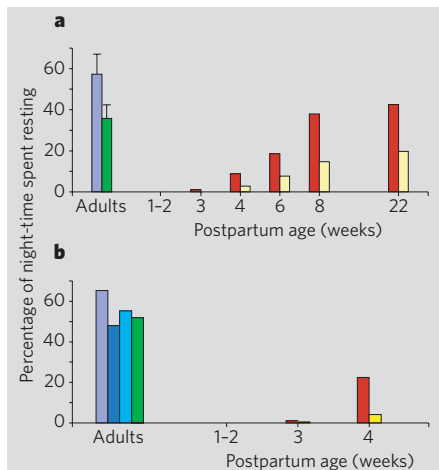


Figure 1 | Duration of rest behaviour in whales and dolphins. a, b, Time spent by killer whales floating at the surface and lying on the bottom of the pool (a) and by bottlenose dolphins floating at the surface (b). The data represent percentages of the night, the main sleep period in captive cetaceans. Data for one non-pregnant orca female and one adult male are the averages (with s.e.m.) of 12 nights¹⁰. All other killer-whale, dolphin and calf data are averages over two consecutive nights. Red, mothers; yellow, calves; blues, adult females without calves; green, adult males.

Administration of cortisol to an adult male and to an adult female dolphin (0.4–0.8 mg per kg, twice a day) and administration of oxytocin — a hormone that increases with parturition and lactation — to another adult female dolphin (0.1 mg per kg, twice a day) produced non-significant increases or no change in the amount of sleep (results not shown).

When the dolphin and killer-whale neonates surfaced to breathe at intervals of 3–30 s, the mother typically continued swimming, forcing the neonate to track her movements and catch up. The non-circular shapes of all but one of the pools housing the dolphins and killer whales, the tendency of the animals to stay within 1–3 m of the walls, and the presence of other animals in the pool (often swimming in opposite directions), as well as frequent breathing in the neonate, all necessitated much turning and surfacing by the postpartum cetaceans. These activities are incompatible with sustained periods of sleep.

We have previously reported that unihemispheric slow waves in the brains of dolphins and beluga whales (*Delphinapterus leucas*) are invariably linked to closure of the contralateral

eye^{5,6}. We found that unilateral or bilateral eye closure was almost never (for less than 0.4% of 24 h) observed in dolphin mothers during the first 2 months postpartum. The calves showed unilateral eye closure rarely (for less than 1.5% of the time) at less than 1 month of age. This amount increased to 16% of the time at 3 months, with the open eye nearly always (95–99% of the time) being directed towards the mother.

We conclude that, although brief periods of slow-wave activity may have occurred, they would have been few; in the calves, they could not have exceeded 30 s before 1 month postpartum. This is surprising as interrupted sleep in humans and rats is known to be largely non-restorative^{7,8}.

The ability to remain active and responsive after birth has several advantages for newborn cetaceans. Motor activity reduces predation and helps to maintain body temperature until their body mass has increased and insulating blubber develops. Surfacing for respiration occurs at shorter intervals in neonates compared with adults and disrupts sleep. The growth of the brain and body and correlated behavioural development in these cetaceans progresses with minimal resting behaviour, in contrast to the pattern seen so far in other animals, from flies⁹ to mammals².

Oleg Lyamin[†], Julia Pryaslova[‡], Valentine Lance[‡], Jerome Siegel^{*}

^{*}University of California, Los Angeles & VA GLAHS, Sepulveda, North Hills, California 91343, USA
e-mail: jsiegel@ucla.edu

[†]Utrish Dolphinarium, Moscow 119071, Russia

[‡]San Diego State University, San Diego, California 92182, USA

- Hoppenbrouwers, T. & Serman, M. B. *Exp. Neurol.* **49**, 822–838 (1975).
- Roffwarg, H. P., Muzio, J. N. & Dement, W. C. *Science* **152**, 604–619 (1966).
- Shaw, P. J., Tononi, G., Greenspan, R. J. & Robinson, D. F. *Nature* **417**, 287–291 (2002).
- Rechtschaffen, A., Gilliland, M. A., Bergmann, B. M. & Winter, J. B. *Science* **221**, 182–184 (1983).
- Lyamin, O. I. et al. *Behav. Brain Res.* **129**, 125–129 (2002).
- Lyamin, O. I., Mukhametov, L. M. & Siegel, J. M. *Arch. Ital. Biol.* **142**, 557–568 (2004).
- Bonnet, M. H. *Physiol. Behav.* **45**, 1049–1055 (1989).
- Rechtschaffen, A. & Bergmann, B. M. *Sleep* **25**, 18–24 (2002).
- Shaw, P. J., Cirelli, C., Greenspan, R. J. & Tononi, G. *Science* **287**, 1834–1837 (2000).
- Flanigan, W. F. *Sleep Res.* **4**, 139 (1975).

Supplementary information accompanies this communication on Nature's website.

Competing financial interests: declared none.
doi:10.1038/435117a

China's environment in a globalizing world

How China and the rest of the world affect each other.

Jianguo Liu and Jared Diamond

China is the world's most populous country and the fourth largest in area. Its economy, already huge, is growing at the fastest rate of any major nation. Its environmental problems are among the most severe of any major country, and are mostly getting worse.

Many Chinese, including its leaders, are aware of these problems and have tried to tackle them. Some things have improved, such as the air quality in Beijing and some other big cities. But such efforts have not matched the forces of environmental destruction, and have not prevented other indicators from further deterioration. The list of problems ranges from air pollution, biodiversity losses, cropland losses, depleted fisheries, desertification, disappearing wetlands, grassland degradation, and increasing frequency and scale of human-induced natural disasters, to invasive species, overgrazing, interrupted river flow, salinization, soil erosion, trash accumulation, and water pollution and shortages. These issues are causing serious economic losses, social conflicts and health costs within China.

China's environmental problems are also spilling over into other countries, while other countries affect China's environment through globalization, pollution and resource exploitation. China is already the largest contributor of sulphur oxides and chlorofluorocarbons to the atmosphere¹; its dust and aerial pollutants are transported eastwards to neighbouring countries and even North America; and it is one of the two leading importers of tropical rainforest timber², making it a driving force behind tropical deforestation. China accounts for 15% of the world fish catch and 33% of global fish and seafood consumption^{3,4}. A factor exacerbating many environmental problems in China is that, as a 'world factory', China exports products but consumes natural resources and leaves pollutants behind. Although China's per capita environmental impact is still far below that of developed

IMAGE
UNAVAILABLE
FOR COPYRIGHT
REASONS

Global problem: pollution and floating rubbish at China's Three Gorges Dam exemplify the challenges for environmental protection across the world.

countries (Table 1), the proportionate increase in total human impact on the world's environments will be enormous if China's per capita impacts catch up with such countries.

After setting out some background information about China, we shall discuss the types of Chinese environmental impacts, their consequences for the Chinese, reciprocal impacts of China and other countries, China's future prognosis, and some recommendations. (Most references and data sources are listed in the Supplementary Information.)

Geography, population, economy and policy

Geography. China's environment is complex (Fig. 1, overleaf). It includes the world's largest and highest plateau, some of the world's

highest mountains, two of the world's longest rivers (the Yangtze and Yellow Rivers), many lakes, a long coastline and a large continental shelf. Its ecosystems range from glaciers and deserts to grasslands, wetlands, tropical rainforests, lakes and oceans (Supplementary Fig. 1). Within those ecosystems lie areas fragile for different reasons: for example, northwestern China's variable rainfall, winds and droughts expose its high-altitude grasslands to dust storms and soil erosion. Conversely, southern China is wet, but heavy rainstorms cause erosion on slopes.

Population. China's population of 1.3 billion people — 20% of the world's total — has more than doubled over the past half century (Supplementary Fig. 2). It is encouraging that the population growth rate has fallen from 2–3% per year between the 1950s and mid-1970s to less than 1% per year in recent years. This is due to a reduction in birth rate, thanks to factors such as the one-child policy. The death rate has remained quite stable for the past 25 years (Supplementary Fig. 3). China's fertility rate in 2003 (1.9 births per woman) (Supplementary Table 1) and population growth rate in 2003 (0.7%) were the fourth lowest among the 15 major countries that we tabulate (Table 1).

But another factor has worked in the opposite direction: the number of China's households grew almost three times as fast as its population during 1985–2000, because average household size decreased from 4.5 to 3.5 people^{5,6}. This alone gave China an extra 80 million households in 2000: more than the total number of households in Russia and Canada combined. All of our comparison countries except Pakistan and perhaps Russia also showed decreasing household size, but China's decrease, and hence its ratio of household number increase to population growth, was the second largest (Table 1). Because smaller households consume more resources per person⁵, China's rapid increase in household number and reduction in household size

have had significant environmental consequences. For instance, while China's household size has been declining, its per capita house floor area has increased more than threefold from the late 1970s to the present (Supplementary Fig. 4).

China is also becoming more urban. From 1952 to 2003, while its total population 'merely' doubled, its proportionate urban population tripled from 13% to 39%. Hence the urban population increased sevenfold to more than half a billion (Supplementary Fig. 2). The number of cities increased fourfold to more than 660 (including more than 170 with at least one million residents), and the areas of existing cities grew hugely.

Economy. China's economy is big, and growing fast (Fig. 2). It ranks third in total gross domestic product (GDP) and has the highest growth rate, of three times the world average, of our 15 comparison countries (Table 1). It is the world's largest producer of steel, cement, aquacultured food and television sets, and is the second-largest producer of electricity and chemical textiles. From 1978 to 2003 its production of steel, cement, chemical fibre and colour TVs increased by 7, 13, 42 and 17,214 times, respectively (Supplementary Fig. 5). It is the largest consumer of fertilizer and accounts for 90% of the global increase in fertilizer use since 1981. As the second-largest producer and consumer of pesticides, China accounts for 14% of the world total and has become a net exporter. Production and consumption of these industrial and agricultural products leads to air, water and land pollution and other forms of environmental damage. But despite China's large total GDP and outputs of these various products, its per capita GDP and outputs are still much lower than those of many other countries — hence they still have a large potential to increase.

With increasing affluence, China's per capita consumption of meat, milk and eggs increased four-, four- and eightfold, respectively, between 1978 and 2002; its egg consumption now equals that of rich nations. This means more agricultural wastes, animal droppings (already four times the output of industrial solid wastes), fish droppings, fish food and fertilizer for aquaculture, tending to increase terrestrial and aquatic pollution.

China's transportation network and number of vehicles have grown explosively (Fig. 3). In 1994, after the number of motor vehicles had increased to six times the 1980 figure, China decided to make car production one of its four 'pillar industries' to stimulate economic growth, with the goal of increasing production (especially of cars) by another factor of four by 2010. This would make China the world's third-largest vehicle manufacturer, after the United States and Japan — with obvious implications for highway expansion at the expense of arable land, greater dependence on imported oil, and the recently improved but still poor air quality in cities such as Beijing.

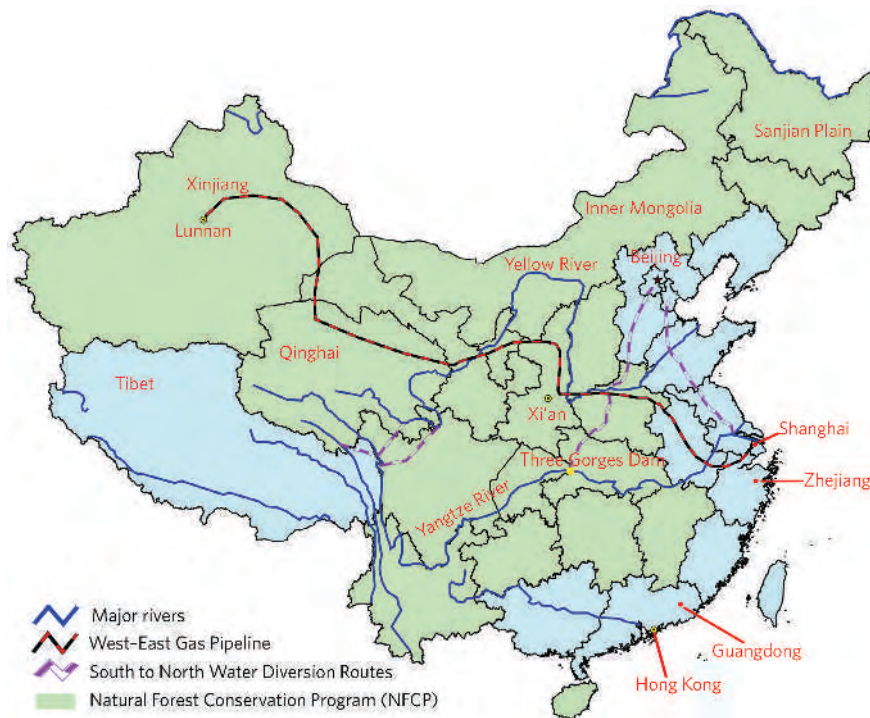


Figure 1 | China. Map of China showing locations of selected projects and places discussed in the text.

Behind these impressive statistics lurks a mixed picture. In sectors of the economy facing strong foreign competition and receiving foreign investment, such as automobile production and fuels, Chinese industry is almost as efficient as that in developed countries. Since 1980, the reduction in China's energy intensity (energy consumption per dollar of GDP) has been unprecedented among developing countries, thanks to energy conservation, phasing-out of inefficient old facilities, adoption of modern technologies and shifts from energy-intensive heavy industry to less intensive light industry and service sectors.

In contrast, much of China's economy — such as coal-mining and cement, paper and chemical production — still rests on outdated, inefficient or polluting technology, and overall industrial energy efficiency is only half that of the developed world (Supplementary Fig. 4). China's paper production consumes more than twice as much water as that in developed nations. Its irrigation relies on inefficient surface methods that waste water, cause eutrophication and wash nutrients out of the soil and sediment into the rivers. China's coal-based production of ammonia, required for fertilizer and textile manufacture, consumes 40–80 times more water than natural-gas-based ammonia production⁷ (because its gas reserves are far from ammonia production centres), although this situation is changing, as discussed below.

Because energy is essential for China's rapid economic development, China is the second-largest energy consumer, after the United States. But China's per capita use of energy in 2001 was only a ninth of that in the United States, and half of the world average. China ranks eighth in that respect among the

15 major countries (Supplementary Table). China leads the world in the production and consumption of coal⁸, with 25% of the world's total. It is the country's primary energy source and the main cause of its air pollution and acid rain, although coal use has declined since the 1950s and has fluctuated in recent years as the use of oil, natural gas and hydroelectric power has increased (Supplementary Fig. 6). In 2003 China overtook Japan to become the second-largest consumer of petroleum after the United States⁹. Although solar and wind power are potentially significant renewable energy sources, hydroelectricity will become more important over the next decade, particularly with the expected completion of the controversial 18.2-gigawatt Three Gorges Dam project in 2009 (Fig. 1).

Natural gas accounts for just 3% of China's energy consumption today. But its use may increase fourfold by 2010 through increases in production from domestic reserves (53.3 trillion cubic feet at the beginning of 2004), and through imports, by pipeline and in the form of liquified natural gas (LNG). The world's longest gas pipeline, the West-to-East Pipeline, began construction in July 2002 to carry gas 3,800 km from the largest reserves in western and north-central China to Shanghai on the east coast (Fig. 1). It will be completed in 2005 (ref. 9). A pilot LNG project is under construction in the economic hotbed of Guangdong Province and will provide 4 billion cubic metres of natural gas annually (Fig. 1).

Another distinctive feature of China's economy is its widely distributed small-scale rural industry: township and village enterprises (TVEs) with an average of six employees (Supplementary Fig. 7). They account for a third of Chinese production and half of its exports but

Table 1 | Population, economy and environmental conditions of China and 14 other major countries*

Country	Population total (millions, 2003)	Annual population growth rate (% 2003)	Ratio of growth in household numbers to population growth (1985–2000)	Average annual GDP growth (% 1999–2003)	Ranking of environmental sustainability index (1–142)** 2002	CO ₂ emission (metric tons per capita, 2000)	Total CO ₂ emission (million metric tons, 2000)	Per capita ecological footprint (global ha per person, 2001)	SO ₂ per populated area (1,000 metric tons per km ² , 2000)
China	1,288	0.7	2.7	8.0	129	2.2	2,780	1.5	2.7
Bangladesh	138	1.7	1.5	5.2	86	0.2	30	0.6	0.7
Brazil	177	1.2	1.9	1.6	20	1.8	310	2.2	0.4
India	1,064	1.5	1.2	5.8	116	1.1	1,120	0.8	1.2
Indonesia	214	1.3	1.8	2.0	100	1.3	270	1.2	0.4
Japan	127	0	6.1	1.3	78	9.3	1,180	4.3	1.0
Malaysia	25	1.9	1.3	4.9	68	6.2	140	3.0	1.6
Mexico	102	1.4	1.9	2.4	92	4.3	420	2.5	1.0
Nigeria	136	2.1	2.7	4.1	133	0.3	40	1.2	0.2
Pakistan	148	2.4	0.4	3.4	112	0.8	110	0.7	0.3
Philippines	82	1.9	1.4	4.3	117	1.0	80	1.2	0.9
Russia	143	–0.4	No data	6.7	72	9.9	1,440	4.4	0.9
Thailand	62	0.6	2.6	4.7	54	3.3	200	1.6	1.1
United States	291	0.9	1.6	3.2	45	19.8	5,590	9.5	1.7
Vietnam	81	1.1	1.5	6.5	94	0.7	55	0.8	0.3
World	6,271	1.2	1.6	2.5	—	4.0	24,210	2.2	1.7

*The most populous countries in the world, with at least 100 million people each, plus the four next most populous countries (Malaysia, Philippines, Thailand and Vietnam) in Southeast Asia.

**1 = most sustainable, 142 = least sustainable, among 142 countries ranked.

contribute disproportionately to pollution^{1,10}. Technology levels in some TVE sectors are advanced, but they are low in other sectors such as brick-making, coal-mining, cement-making, paper production, pesticide and fertilizer manufacturing, coking and metal-casting, which consume more resources and produce more pollution than larger state-owned enterprises.

Policy. China's leaders once believed that humans could and should conquer nature, and that only capitalist societies suffered from environmental damage¹¹. Such thinking began to change in 1972, when China sent a delegation to the First United Nations Conference on the Human Environment¹¹. In 1973 the government's Leading Group for Environmental Protection was established, which evolved in 1988 into the National Environmental Protection Agency, and in 1998 became the State Environmental Protection Administration (SEPA)¹². China declared environmental protection a basic national principle in 1983, laid out a broad strategy to achieve sustainable development in 1994, and in 1996 developed its first five-year plan on environmental protection¹². In 2003, the government proposed a new development concept emphasizing humanism and attempting to achieve sustainable development and harmony between man and nature, as well as coordinated socio-economic progress among various regions and with foreign countries¹³. China has also participated in international treaties such as the Convention on Biological Diversity and the UN Millennium Development Goals, which include poverty alleviation, environmental protection and sustainable development. More than 100 environmental policies, laws and regulations have been passed. These seem

excellent on paper, but putting them into practice is not easy. In reality, although there has been much effort to control environmental degradation, economic development often takes priority at the local level and is still the main criterion for judging government officials' performance.

Environmental impacts

There was large-scale deforestation in China several thousand years ago. Following the Second World War and the Chinese Civil War, the peace of 1949 brought more deforestation, overgrazing and soil erosion. The Great Leap Forward in 1958–1960 saw a dramatic increase in the number of factories — there was a fourfold increase in 1957–1959 alone — along with pollution and more deforestation, to obtain the fuel for inefficient backyard steel production. From the 1960s until the mid-1970s, pollution grew, as many factories were relocated to the interior from coastal areas considered militarily vulnerable. Since economic reform began in 1978, environmental degradation has continued to accelerate^{10,14}, largely due to rapid industrialization, including TVEs.

China faces greater environmental challenges than other major countries. Of the 142 countries for which environmental sustainability was evaluated, China ranked 129th, higher only than Nigeria among our 15 comparison countries (Table 1). In per capita ecological footprint (a measure of human natural resource consumption and waste output), China is below the world average (Supplementary Table), but its total ecological footprint is the second largest in the world after the United States, owing to its population size.

China's environmental problems can be

summarized under five categories: air, land, fresh water, oceans and biodiversity.

Air. China's air quality is generally low. Three out of four city dwellers live below China's air-quality standard¹⁵. Acid rain fell on a quarter of its cities for more than 60% of rainy days per year in the 1990s and now affects a quarter of China's area, making it among the world's most severely affected countries⁸.

A major cause of these problems is the increasing output of industrial waste gases (Supplementary Fig. 8). After declining or levelling off in 1998, emissions of SO₂ and possibly of dust and industrial soot resumed climbing in 2003. In 2000 China led the world in SO₂ emissions (Table 1) and ranked third for NO_x emissions in populated areas among the major countries (Supplementary Table).

On the other hand, several air-quality indicators have shown positive signs. More industries are achieving emission standards. Among the 47 key cities for environmental protection, 11 and 29 have exceeded the national air-quality standards for SO₂ and particulate concentrations, respectively, including Beijing¹⁵.

Land. Soil erosion affects 19% of China's land area, one of the highest figures for any country¹⁰. Erosion is especially devastating on the Loess Plateau on the middle stretch of the Yellow River, which is about 70% eroded, and increasingly on the Yangtze River, whose sediment discharge from erosion exceeds the combined discharges of the Nile and Amazon, the world's two longest rivers. By filling up rivers (as well as reservoirs and lakes), sediment has shortened China's navigable river channels by 56% between 1949 and 1990, and has restricted the size of ships that can use them. Soil quality and fertility, as well as soil

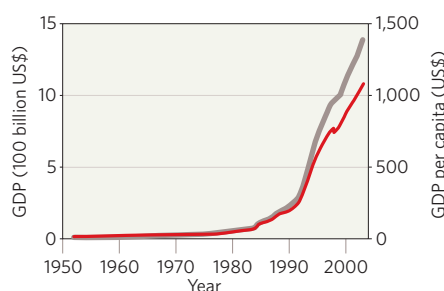


Figure 2 | Chinese gross domestic product (GDP). Growth of national (brown line) and per capita (red line) GDP.

quantity, have declined, in part due to long-term fertilizer use plus pesticide-related declines in soil-renewing earthworms. Salinization has affected 9% of China's lands, mainly due to poor design and management of irrigation systems. This is one environmental problem that government programmes have made good progress in combating and starting to reverse. Desertification, due to overgrazing and land reclamation for agriculture, has affected more than a quarter of China, especially in Qinghai Province and the Inner Mongolia Autonomous Region.

All of these soil problems have joined urbanization and land appropriation for mining, forestry and aquaculture in reducing China's cropland. This threatens food security¹⁶, because while cropland area has been declining, population and per capita food consumption have been increasing, and the area of cultivatable land is limited. Between 1991 and 2000, cropland declined to the point where there is now only 0.1 ha per person, barely half of the world's average. Unrecycled and unused industrial waste and domestic trash are dumped into open fields around most cities, polluting soil and taking over or damaging 100,000 km² of cropland¹⁴. Industrial solid-waste production has risen, but waste release is declining because of increased recycling (Supplementary Fig. 9).

China is one of the world's most forest-deficient countries, with only 0.1 ha of forest per person, compared with a world average of 0.6 ha. Forests cover only 18% of China's land area, compared with 64% of Japan's and 30% on average (Supplementary Table). Although government programmes have increased the area of single-species tree plantations and thereby the total forested area (Supplementary Fig. 10), natural forests, especially old growth, have shrunk. Deforestation is a major cause of soil erosion and flooding in China. The 1998 floods that affected 240 million people shocked the government into action, including the banning of any further logging of natural forests in upper and middle reaches of watersheds of major rivers such as the Yangtze and Yellow Rivers.

The other most serious forms of land degradation are the destruction of grasslands and wetlands. China is second only to Australia

in the extent of its natural grasslands¹, which cover 40% of its area¹⁰, mainly in the drier northwest. However, per capita grassland area is less than half of the world's average. Grasslands have been declining at approximately 15,000 km² a year since the early 1980s. Furthermore, grasslands have been severely degraded by overgrazing, climate change, and mining and other types of development; 90% of China's grasslands are now considered degraded. Grass production per hectare has decreased about 40% since the 1950s, and weeds and poisonous grasses have thrived at the expense of high-quality species. Grassland degradation has implications beyond its usefulness to China's farmers, because the grasslands of the Tibetan Plateau contain the headwaters for the major rivers of India, Pakistan, Bangladesh, Thailand, Laos, Cambodia and Vietnam, as well as of China (Supplementary Fig. 1).

There are approximately 660,000 km² of wetlands in China, about 10% of the world's total. However, wetlands have been decreasing in area through conversion to cropland and other uses. Three-fifths of the swamps in the Sanjian Plain in the northeast, the area with China's largest freshwater swamps, has already been drained to become farmland. At the present rate the rest will disappear within 20 years¹⁰. As a result, natural wetlands account for only 3.8% of China's territory, less than the global level of 6.0%. Wetland function has also declined, with greater water-level fluctuations and reduced capacity to mitigate floods and to store water. Wetlands face other major threats, including increased pollution, insufficient funding, and ineffective laws and regulations.

Fresh water. Water quality in most Chinese rivers and groundwater sources is poor and declining, owing to industrial and municipal wastewater discharges, plus agricultural and aquacultural run-offs of fertilizers, pesticides and manure, causing widespread eutrophication⁷. The amount of waste water discharged has increased steadily (Supplementary Fig. 11). About 75% of lakes are polluted. The Guanting Reservoir in Beijing was declared unfit for drinking in 1997. The percentage of industrial waste water treated has been increasing, but only 20% of domestic waste water is treated, compared with 80% in the developed world.

Shortages and waste exacerbate China's water problems. China's per capita quantity of fresh water is only a quarter of the world average. Water resources are spread unevenly, with northern China having only one third of the per capita quantity of southern China. This underlying water shortage, plus wasteful use, causes over 100 cities to suffer from severe shortages and even halts industrial production. Of the water required for cities and for irrigation, two-thirds depends on ground water pumped from wells tapping aquifers. However, those aquifers are becoming depleted, letting sea water enter them in most

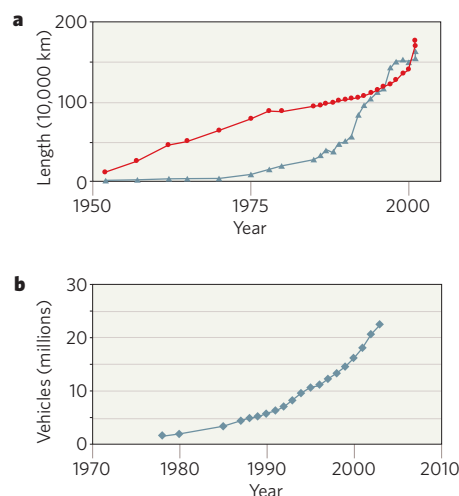


Figure 3 | Transport in China. a, Length of highways (red line) and number of civic aviation routes (grey line). b, Number of vehicles.

coastal areas, and causing subsidence in some cities as the aquifers are drawn down. China already has the world's worst cessation of river flows, and this is increasing because people continue to draw water from rivers. There were flow stoppages on the lower Yellow River in 20 of the years between 1972 and 1997, and the number of days without any flow increased from 90 days in the 1980s to an astonishing 230 in 1997 (ref. 10).

Pollution and overfishing are degrading freshwater fisheries because fish consumption is rising steeply. Per capita fish consumption has increased nearly fivefold in the past 25 years¹⁷, and there is a growing export of fish, molluscs and other aquatic species. As a result, the white sturgeon has been pushed to the brink of extinction, previously abundant fish species such as the yellow croaker and hairtail must be imported, the catch of wild fish in the Yangtze River has declined by 75%, and that river had to be closed to fishing for the first time in 2003 to protect fishery resources from collapse. To meet demand for fish products, production of aquacultured freshwater fish has increased steeply (Fig. 4).

Oceans. China has a sea area of 3 million km² and has jurisdiction over the vast continental shelves and exclusive economic zones up to 200 nautical miles off its coasts. Almost all coastal seas are polluted¹, mainly by pollutants from the land, plus oil spills and other marine activities. In 2004 the State Oceanic Administration recorded 867 main outlets discharging pollutants into the sea. In 2003 alone, 20 of those outlets discharged approximately 880 million tonnes of sewage water, containing 1.3 million tonnes of pollutants, including toxic substances such as lead, cadmium and arsenic. On average, there are 90 red tides in China's seas each year, up from only one every five years in the 1960s (Fig. 5). Pollution and overfishing have hit fishery stocks. Natural harvests have significantly declined — the formerly robust Bohai prawn harvest has

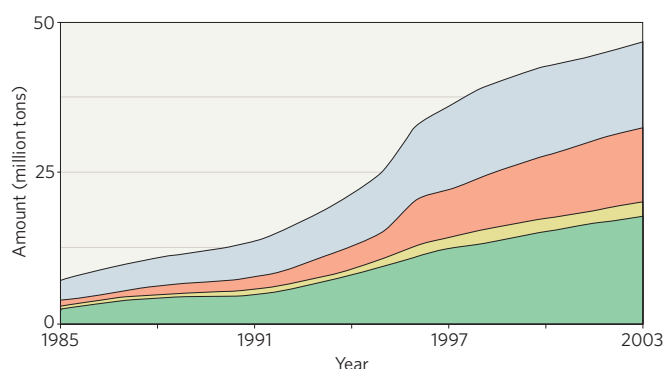


Figure 4 | Chinese aquatic production. Mass of cultured freshwater (green), wild freshwater (yellow), cultured marine (orange) and wild marine (grey) products.

dropped by 90% — and production of aquacultured seafood has increased (Fig. 4). China's area of mangrove declined by 73% from the early 1950s to 2002.

Biodiversity. China has more than 10% of the world's vascular plant and terrestrial vertebrate species¹⁸. However, 15–20% of China's species — including the giant panda — are now endangered, largely by human activities¹. Many distinctive rare animals and plants, such as Chinese alligators, are at risk of extinction. To protect biodiversity, the Chinese government had set up almost 2,000 nature reserves by the end of 2003, mostly within the past 20 years (Fig. 6), plus many zoos, museums, botanical gardens, wildlife breeding centres, and gene and cell banks. The reserves cover 14.4% of China's territory, a percentage higher than the world average and than the percentages of most developed countries. Nevertheless, these reserves must be better managed and more strategically important reserves are needed¹⁸.

The flip side of these declines in native species has been rises in both terrestrial and aquatic invasive species: more than 400 by 2004. Examples include ragweed (a plant native to North America), water hyacinth and Amazonian snails¹⁹. Some of those invaders have become pests and weeds, inflicting heavy economic damage on Chinese agriculture, aquaculture, forestry and livestock production: almost US\$14.5 billion (1.4% of China's GDP in 2000) in 2000 alone. Most invasive species were brought into China, intentionally or unintentionally, by international trade and other activities. In Shanghai harbour alone, between 1986 and 1990, almost 200 foreign weed species were found in imported materials carried by 349 ships from 30 countries.

Consequences for China's people

China's environmental degradation is harmful not only to its earthworms and yellow croakers, but also to its people. The consequences for Chinese people can be partitioned into socioeconomic losses, health costs, and the effects of more frequent and damaging natural disasters.

Socioeconomic losses. Starting with small examples and proceeding to larger ones: \$72 million per year is being spent to curb

of crops and forests due to acid rain amount to about \$730 million per year²². More serious is the \$6 billion cost of the 'green wall' of trees being built to shield Beijing against sand and dust, the annual direct losses due to desertification (\$7 billion), and the \$7 billion per year in losses created by several major alien species other than alligator weed. Even bigger numbers are the one-off cost of the 1996 floods (\$27 billion, but still cheaper than the 1998 floods) and the annual losses due to water and air pollution (\$54 billion)^{7,10}.

The losses from pollution and ecological damage ranged from 7% to 20% of GDP every year in the past two decades²³. Besides heavy economic losses, pollution and resource competition have triggered numerous social clashes in China, including 18 conflicts over forest resource management in southwestern China compiled by the Food and Agriculture Organization of the United Nations in 2001. Similarly, water shortages in the Yellow River have triggered 'water wars' between people on the river's upper and lower reaches, between people on opposite sides of the river, and between backers of industrial, agricultural and ecological needs.

Health costs. Environmental pollution imposes further costs through its impact on human health. From 1996 to 2001, China's spending on public health increased by 80%, or more than 13% per year (from \$35 billion in 1996 to \$63 billion in 2001)²⁴, in part to cope with environmental problems. About 300,000 deaths per year are attributed to air pollution⁷. Average blood lead levels in Chinese city dwellers are nearly double those considered to be dangerously high and to endanger children's mental development. The risk of respiratory disease increases with the outdoor concentration of total suspended particles. Even short-term exposure to air pollution can result in low infant weight and increased morbidity and mortality²⁵.

Natural disasters. China is noted for the frequency, number, extent and impact of its natural disasters. Human actions have made some of these more frequent, especially dust storms, landslides, droughts and floods¹⁰. Overgrazing, erosion, grassland degradation,

the spread of a single weed²⁰, the alligator weed, introduced from Brazil as pig forage. It has spread to infest gardens, sweet-potato fields, and citrus groves. Also relatively cheap is the annual loss of \$250 million arising from factory closures due to water shortages in a single city, Xian¹⁰. Sand-storm damage costs about \$540 million per year²¹, and losses

desertification and partly human-caused droughts have led to more frequent, and more severe, dust storms. From AD 300 to 1949, dust storms struck northwestern China on average once every 31 years; since 1990 there has been one almost every year. The huge dust storm of 5 May 1993 killed a hundred people. Recent increases in droughts are believed to be due to deforestation that has interrupted the water cycle, and perhaps also due to the decrease in surface water resulting from draining and overuse of lakes and wetlands. Droughts damage about 160,000 km² of cropland each year, double the area damaged in the 1950s. Flooding has greatly increased because of deforestation; the 1996 and 1998 floods were the worst in recent memory. Alternating droughts and floods have become more frequent and are more damaging than either disaster alone, because droughts destroy vegetation, and then flooding of bare ground produces worse erosion.

How China and the world affect each other

China and the rest of the world have become closely interconnected. China's large territory and population guarantee environmental impacts on the rest of the world. The rest of the world increases these impacts by means of the trade and investment that fuel China's rapid economic growth. Although international trade was negligible before 1980 (Fig. 7a), and although foreign investment in China was negligible as recently as 1991 (Fig. 7b), both have recently accelerated almost exponentially. There was a 40-fold increase in international trade between 1978 and 2003.

Since 2002, China has overtaken the United States to receive the most foreign investment annually of any country (Supplementary Table). The Chinese government has encouraged foreign investment through the development of 'special economic zones' in which foreign investors receive preferential tax and tariff treatment. Environmental impacts of foreign investment and international trade may be either a positive or negative^{26,27}, as we will now show.

Beneficial and harmful imports. Much of the products, technologies, knowledge and financial support imported into China is environmentally benign or strongly beneficial. Between 1992 and 2004 the World Bank provided more than \$22 billion to China, of which approximately 10% was used for environment-related projects. Many of the imported raw materials and products help China reduce its consumption of domestic natural resources and its pollutant discharge. For example, agricultural imports let China decrease its use of fertilizers, pesticides, water and low-productivity cropland; and oil and natural-gas imports let China reduce pollution from burning coal. Since 1993, China's oil consumption has exceeded its oil production, and the gap is widening⁹. From 1980 to 2002, the value of China's imported primary

goods increased from \$7 billion to \$49 billion.

On the other hand, some imports are unequivocally harmful to China's environment. Along with the invasive species mentioned earlier, another example is imported garbage. Some developed countries export untreated garbage to China, including waste containing toxic chemicals. In addition, China's expanding manufacturing economy accepts garbage and scrap that could be a cheap source of recoverable raw materials. As just one example, in September 2002 a customs office in Zhejiang Province recorded a 360-tonne shipment of electronic garbage from the United States, consisting of scrap electronic equipment and parts such as broken or obsolete TV sets, computer monitors, photocopiers and keyboards. Statistics on the total amount of such garbage imported are incomplete, but estimates show an increase in direct imports from 1 million to 11 million tonnes from 1990 to 1997 (ref. 28), and garbage shipped via Hong Kong (Fig. 1) also increased from 2.1 million to over 2.7 million tonnes per year from 1998 to 2002. Although some people view importing harmful garbage as part of normal international trade, the Chinese government prohibits it and has been trying to stop it.

Even worse than garbage, while many foreign companies have helped China's environment by transferring advanced technology to China, others have hurt it by transferring pollution-intensive industries (PIIs), including technologies illegal in the country of origin. As of 1995, China was home to an estimated 16,998 PII firms with a combined industrial product of about \$50 billion²⁸. For financial and various other reasons, it has often been impossible for China to adopt the advanced technology standards of developed nations, which in turn profit and gain competitive advantage by exporting outdated or even illegal technologies. Many Chinese officials and economists believe that PIIs benefit China by raising economic efficiency and reducing pollution in the long run. But PIIs cause severe damage to the environment, as well as to human health and socioeconomic well-being, and some of the damage, such as biodiversity losses, is irreversible.

Exports causing damage at home. Export trade is a major cause of China's increasing pollution, because products go abroad but pollutants stay behind. Most of China's exports are primary goods or manufactured products that create heavy pollution and require intensive resource uses. For instance, from 1989 to 2002 the value of goods exported by heavily polluting TVEs increased 31-fold, including a 22-fold increase in textiles and an 18-fold increase in food²⁹.

Invasive species exported. China's high native biodiversity means that China exports many invasive species. The three best-known pests of North American tree populations — the chestnut blight, the misnamed 'Dutch'

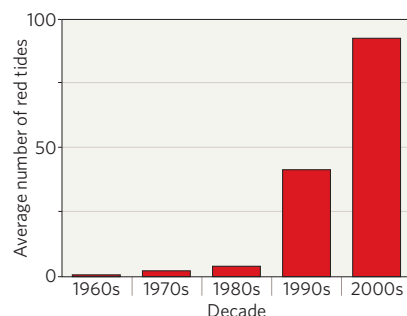


Figure 5 | Average annual number of red tides in Chinese seas.

elm disease, and the Asian long-horned beetle — originated in China or somewhere nearby in East Asia¹⁹. China's grass carp is established in rivers and lakes of 45 US states, where it competes with native fish species and changes the plant, plankton and invertebrate communities.

Exports in the atmosphere. China became the world's largest producer and consumer of ozone-depleting gases, such as chlorofluorocarbons, after developed countries phased them out in 1995 (ref. 1). China already leads the world in the production of sulphur oxides, with an output double that of the United States. China's per capita production of CO₂ and NO_x is far below that of rich countries, or for CO₂ even below Mexico, Russia and Thailand (Table 1 and Supplementary Table). But China's huge population is still the second-largest contributor of CO₂, emitting approximately 12% of the world's total.

Aerial particles from China also affect the regional and global atmosphere. Propelled eastwards by prevailing winds, pollutant-laden dust, sand and soil from China's deserts, degraded pastures and fallow farmland blow to Korea, Japan, Pacific islands and across the Pacific within a week to the United States and Canada³⁰. The aerial particles result from China's coal-burning economy, overgrazing and soil erosion. Together with affected countries and the international community, China has been trying to reduce aerial particles as well as greenhouse gases and ozone-depleting substances.

Exported deforestation. China ranks third in the world in timber consumption¹. Because wood provides almost all the raw material for the paper and pulp industry, and also panels and lumber for construction, there is a growing gap between China's demand for wood products and its domestic supply, especially since the national logging ban that followed the floods of 1998. China's wood imports, both from tropical and temperate countries, have increased sixfold since the ban³¹. As an importer of tropical lumber, China now stands second only to Japan, which it is rapidly overtaking. With China's entrance into the World Trade Organization (WTO), timber imports are expected to increase, because tariffs on wood products are about to be reduced from a rate of 15–20% to 2–3%. In effect, this means

that China, like Japan, will be conserving its forests by exporting deforestation³¹, already at or close to devastating levels in several countries, including Malaysia, Papua New Guinea and Australia.

The future

What does the future hold for China? Environmental problems are accelerating, and attempted solutions are accelerating, but which horse will win the race?

Generalized dangers. A pessimist will note many dangers already at work in China. Economic growth, rather than environmental protection or sustainability, is still China's priority in practice. Despite a fall in population growth rate, the number of Chinese is projected to reach almost 1.5 billion by 2030. The projected drop in household size to 2.2 people⁶ by the year 2030 alone would add over 250 million new households — more than the total in the entire Western Hemisphere in 2000 — even if China's population size remained constant.

Public environmental awareness is low, in part because China's investment in education is less than half that of developed countries as a proportion of gross national production. Despite holding 20% of the world's population, China's educational funding accounts for only 1% of world investment. Most parents cannot afford to send their children to university, because one year's tuition would consume the average salary of one city worker or three rural workers.

Chinese environmental laws and regulations were written largely piecemeal, lack effective implementation and evaluation of long-term consequences, and need a systems approach. Prices for important environmental resources are set so low as to encourage waste: one could buy 10–100 tonnes of Yellow River water for use in irrigation for the cost of a small bottle of spring water¹⁰. Land is owned by the government, but may be used by many different peasants within a relatively short period, so peasants lack incentives to make long-term investments in their land or to take care of it.

Specific dangers. The Chinese environment also faces many specific dangers. The number of cars is rising, and croplands and natural wetlands are disappearing. The harmful consequences of this will accumulate. With rising affluence, and hence meat and fish consumption, environmental problems from meat production and aquaculture, such as pollution from animal and fish droppings and eutrophication from uneaten fish food, will increase. Already, China is the world's largest producer of aquaculture-grown food, and is the sole country in which aquaculture provides more fish and aquatic foods than wild fisheries.

China is hosting the world's three biggest development projects (Fig. 1), all of which are expected to cause severe environmental problems. The Three Gorges Dam on the Yangtze River — the world's largest dam, begun in

1993 and projected for completion in 2009 — aims to provide electricity, flood control and improved navigation at a cost of \$30 billion, social costs of uprooting millions of people, and environmental costs associated with landslides, water pollution, soil erosion, biodiversity losses and the disruption of the ecosystem of the world's third-longest river³². Still more expensive is the South-to-North Water Diversion Project, which began in 2002 but is not scheduled for completion until around 2050. It is projected to cost \$59 billion, to spread pollution, and to cause water imbalance in the Yangtze. Even that project will be exceeded by the ongoing development of western China, comprising over half of the country's land area and viewed by China's leaders as the key to national development.

Increased world impact. Potentially more important than all of these other impacts is a further consequence of China's having the world's largest population and fastest-growing economy. Total production or consumption is the product of population size times per capita production or consumption rate. China's total production and consumption are already high, because of its huge population, despite its per capita rates still being very low. For instance, the per capita consumption rate of four major industrial metals (steel, aluminum, copper and lead) is only 9% of that of the leading industrial countries. But China is rapidly becoming a developed-world economy. If China's per capita consumption rates do reach such levels, and even if populations, production and consumption rates everywhere else remained unchanged, those rate increases alone would translate into a 94% increase in total world production or consumption in industrial metals, and a 106% increase in the case of oil. In other words, China's achievement of developed-world consumption standards will approximately double the world's human resource use and environmental impact. But it is doubtful whether even the current human resource use and impact on the world can be sustained. Something has to give, or change. This is why China's environmental problems are the world's.

Hopeful signs. There are also important sources of optimism. China is increasingly assuming responsibilities on the world stage by participating in environmental treaties. Many environmental laws, policies and regulations are being developed or improved. The Chinese public's environmental awareness is rising. China has been pushing hard for cleaner production and sustainable development. Some environmental and product standards have reached developed-world levels. Energy intensity is declining. Technologies for production and for treating environmental waste are improving.

China has promoted the use of ecological principles in production and pollution control, such as ecological agriculture and some traditional environmentally friendly technologies.

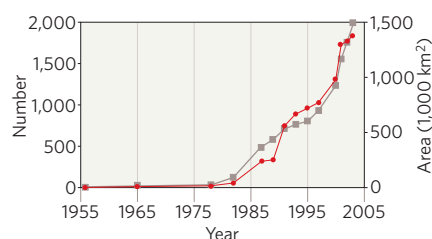


Figure 6 | Chinese nature reserves. Shown by number (brown line) and area (red line).

For example, the southern Chinese practice of raising fish in irrigated rice fields recycles fish droppings as fertilizer, increases rice production, uses fish to control insect pests and weeds, decreases herbicide, pesticide and synthetic fertilizer use, and yields more dietary protein and carbohydrate, without increasing environmental damage.

Both WTO membership and the impending 2008 Olympic Games in Beijing have made the Chinese government pay more attention to environmental problems. To reduce air pollution in Beijing, the city government ordered that vehicles be converted to allow the use of natural gas and liquefied petroleum gas. China has phased out leaded petrol in little more than a year, something that took Europe and America many years to achieve. New cars must meet the exacting emissions standards prevailing in Europe.

Also encouraging is the 1998 ban on logging and the start of the Natural Forest Conservation Program (Fig. 1) to reduce the risk of further flooding³³. Since 1990, China has combated desertification on 24,000 km² of land by reforestation and fixation of sand dunes¹⁰. The Grain-to-Green programme, begun in 2000, gives grain and cash subsidies to farmers who convert cropland to forest or grassland, and is reducing the use of environmentally sensitive steep hillsides for agriculture. By the end of 2003, 79,000 km² of cropland had been returned to forest or grassland³⁴. By the end of this programme in 2010, approximately 130,000 km² of cropland are expected to be converted³⁵, making it one of the largest conservation programmes in the world. China is also designing and adopting a green accounting system that includes environmental costs in the calculation of gross domestic product (or Green GDP).

Recommendations and outlook

How can China turn its environmental trend from deterioration to improvement? Many specific recommendations follow directly from our review. For example, China could import technologies for decreasing fertilizer and pesticide use, reducing motor-vehicle exhaust pollution, improving efficiency of paper and ammonia production and irrigation systems, treating waste water, conserving water and other resources, promoting the use of cleaner energy, and stopping draining of wetlands. We also offer six broader sets of recommendations:

1. The impressive body of environmental laws and regulations that exists largely on paper should be implemented and enforced. Because some governmental officials have interests in companies that damage the environment, it is hard for them to enforce environmental policies. To avoid conflicts of interest, regulation of environmental resources should be transferred to the SEPA from agencies responsible for developing those resources. The SEPA should have the power to close down heavy polluters, because many local officials protect polluters to boost GDP, the main criterion for their promotions. Selection and promotion of government officials should consider environmental protection as well as economic development. The relatively small number of environment enforcement officials should be increased and they should be better trained.

Lack of enforcement is also due to lack of funding. China has a lower GDP than Japan and the United States (Supplementary Table) but more serious environmental problems, so it needs proportionally higher environmental investment. Hence China's budget for environmental protection should rise from its current 1.2% of GDP to rich-nation levels (1.5% in Europe and Japan, and 2% in the United States) or higher. A high investment would make sense on economic grounds alone, by eliminating much of the losses caused by environmental damage.

2. As China moves towards a more market-based economy, more market tools should be applied to environmental issues. Possible examples include: eliminating subsidies for environmentally damaging industries, such as coal; setting fair prices for ecosystem services that are now grossly underpriced, such as water; enhancing emissions trading to reduce pollution; imposing more environmental taxes, such as a higher consumption tax on cars; compensating residents in and around nature reserves, such as those for the endangered giant pandas; and incorporating direct and indirect environmental costs (such as pollution) as well as values of ecosystem services (such as of wetlands) into accounting from local to national levels.

3. Focus attention not only on population size, whose growth has already slowed, but also on household number, size and consumption⁵. The government should provide incentives for sharing household resources.

Two major factors in the dramatic increase in household numbers and reduction in household size are divorces and declines in the number of households where several generations live under one roof. Many older people now live alone, rather than with their children and grandchildren. Divorces have increased sharply owing to simplified divorce procedures and wider societal acceptance of divorce. In 2004, more than 1.6 million couples filed for divorce, up 21% from 2003. Divorces hurt the environment because they double the number

of households and reduce the household size, increasing per capita resource consumption and waste. Government-supported mediation, counselling and a mandatory waiting period of one month or more for divorce would help people to think more seriously about divorce. Incentives should be provided to encourage sharing of resources through schemes such as co-housing (conceived in Denmark) and eco-villages (founded in the United States and Russia). These provide not only socioeconomic benefits to co-habitants, but also help to increase the efficiency of resource use and reduce per capita ecological footprints.

4. Investment in education should be increased significantly. Besides ameliorating China's environmental problems by increasing environmental awareness and decreasing human fertility, educational investments would yield economic benefits by upgrading the skills of China's work force. Better elementary and high-school education would also help more children in biodiversity-rich and environmentally fragile regions, such as western China, to go to college and reduce human pressure on sensitive ecosystems, because college graduates have better opportunities to find jobs and settle down elsewhere.

5. More effective measures should be taken to conserve biodiversity. Polluted air and water can be cleaned up, but lost species and genetic materials cannot be restored. Furthermore, biodiversity offers goods and services essential for human survival, including clean, nutritious food, water and air purification, oxygen generation, mitigation of climate change, pollination of crops and many other plants, control of crop pests, and carbon storage. For example, the naturally sterile male wild-rice variety discovered in China in 1970 has made high-yield hybrid rice possible, and with it the second green revolution.

6. Other countries can, and should, help China to protect its environment. Importing countries contribute to China's pollution. Per capita resource consumption and pollutant outputs are still much lower in China than in developed countries, so China has the moral right, as well as the power, to develop. But the resulting environmental impacts would extend beyond China's borders, making it in other countries' interests to help China. One way would be to support Chinese environmental non-governmental organizations (NGOs), because in China, as elsewhere, those problems exceed governments' capacity to solve them unaided. China has more than 2,000 fledgling environmental NGOs, but most are small, poorly funded, isolated and in need of help. Together with the Chinese government, the international community could help NGOs to increase the public's environmental awareness, contribute to governmental policy, and monitor policy implementation. Other possibilities include: training environmental planners and managers; sharing methods for conflict resolution; transferring environmen-

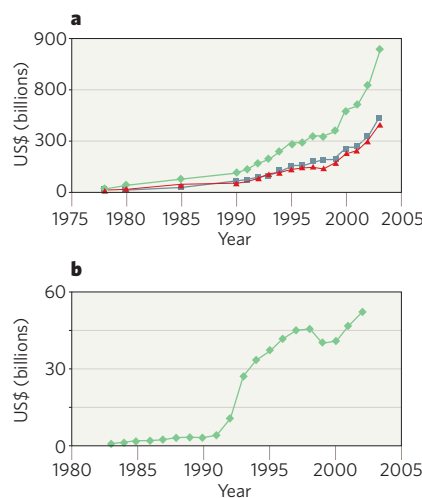


Figure 7 | Trade and foreign investment in China. **a**, Imports (red line), exports (grey line) and total imports and exports (green line); **b**, Foreign direct investment.

tally benign technologies, such as ones for cleaner manufacturing, water conservation and waste treatment; and transferring high-efficiency technologies, which would yield the additional advantage of reducing the already growing competition between China and other countries for energy and for other global resources.

How will it all end up? China is lurching between accelerating environmental damage and accelerating environmental protection. Its large population and booming economy mean that China's lurches carry more momentum than those of other countries. In the past two decades, China has created an economic miracle. We hope that, over the next two decades, China can also create an environmental miracle and set a good example for other nations to achieve both socioeconomic and environmental sustainability. The outcome will affect not just China, but the entire world.

Jianguo Liu is in the Center for Systems Integration and Sustainability, Department of Fisheries and Wildlife, Michigan State University, East Lansing, Michigan 48824, USA. He is also guest professor at the Research Center for Eco-Environmental Sciences, Chinese Academy of Sciences, Beijing, and specially appointed professor, Fudan University, Shanghai, China. e-mail: jliu@panda.msu.edu

Jared Diamond is in the Department of Geography, Bunche Hall, University of California, Los Angeles, California 90095-1524, USA. e-mail: jdiamond@geog.ucla.edu

1. The World Bank *China: Air, Land, and Water* (World Bank, Washington DC, 2001).
2. Adams, M. & Castano, J. <http://www.itto.or.jp/live/contents/download/tfu/TFU.2001.01.English.pdf> (ITTO Newsletter, 2001).
3. United Nations Food and Agriculture Organization <http://faostat.fao.org/faostat> (2004).
4. Pauly, D. *et al. Science* **302**, 1359–1361 (2003).
5. Liu, J., Daily, G., Ehrlich, P. & Luck, G. *Nature* **421**, 530–533 (2003).

6. United Nations Centre for Human Settlements *Cities in a Globalizing World: Global Report on Human Settlements 2001* (Earthscan, London, 2001).
7. Johnson, T. M., Liu, F. & Newfarmer, R. *Clear Water, Blue Skies: China's Environment in the New Century* (World Bank, Washington DC, 1997).
8. Feng, Z., Miao, H., Zhang, F. & Huang, Y. *J. Env. Sci.* **14**, 227–233 (2002).
9. US Energy Information Administration <http://www.eia.doe.gov/emeu/cabs/china.html> (2004).
10. Department of Nature Conservancy (SEPA) *Report on China's Ecological Issues* (in Chinese) (China Environmental Science Press, Beijing, 1999).
11. Qu, G. *Dream and Expectation: Past and Future of China's Environment* (in Chinese) (China Environmental Science Press, Beijing, 2000).
12. Editorial Board *Environmental Protection in China* (in Chinese) (China Environmental Science Press, Beijing, 2000).
13. Wen, J. http://www.sociology.cass.net.cn/shxw/shgz/shgz6/t20040405_1994.htm (in Chinese) (2004).
14. Xie, Z. in *Greening Strategy* (in Chinese) (eds. Lee, Z. & Zhou, G.) 44–46 (Qingdao Press, Qingdao, 1997).
15. State Environmental Protection Administration of China (in Chinese) <http://www.sepa.gov.cn/eic/652466692596695040/20040602/1050958.shtml> (2004).
16. Brown, L. R. *Who Will Feed China?* (Norton, New York, 1995).
17. United Nations Food and Agriculture Organization <http://www.fao.org/News/2001/010702-e.htm> (2001).
18. Liu, J. *et al. Science* **300**, 1240–1241 (2003).
19. Yan, X., Li, Z., Gregg, W. P. & Li, D. *Biodiv. Conserv.* **10**, 1317–1341 (2001).
20. China Daily <http://test.china.org.cn/english/2003/Apr/61832.htm> (2003).
21. Editorial Board *Atlas of Population, Environment, and Sustainable Development of the People's Republic of China* (in Chinese) (Science Press, Beijing, 2001).
22. Yang, L., Stulen, I., De Kok, L. J. & Zheng, Y. *Phyton* **42**, 255–264 (2002).
23. Guo, X. in *China Environment and Development Review Vol. 2* (in Chinese) (Editorial Board) 53–71 (Social Science Documentation Publishing House, Beijing, 2004).
24. National Bureau of Statistics of China *China Statistical Yearbook 2000* (in Chinese) (Beijing, 2001).
25. Xu, X., Wang, L. & Niu, T. *Ecosyst. Health* **4**, 199–209 (1998).
26. Chai, J. C. H. *Sust. Dev.* **10**, 25–35 (2002).
27. Dean, J. M. *Can. J. Econ.* **35**, 819–842 (2002).
28. Ye, R. *et al. Environment and Trade* (in Chinese) (China Environmental Science Press, Beijing, 2001).
29. National Bureau of Statistics of China *China Statistical Yearbook 2003* (in Chinese) (Beijing, 2004).
30. Holden, C. *Science* **294**, 2469 (2001).
31. Global Wood Trade Network <http://www.globalwood.org/news/b091202.htm> (2002).
32. Dai, Q. in *The River Dragon Has Come! The Three Gorges Dam and the Fate of China's Yangtze River and Its People* (ed. Thibodeau, J. G. & Williams, P. B.) pp. 3–17 (Sharpe, Armonk, New York, 1998).
33. Zhang, P., Zhou, X. & Wang, F. *Introduction to Natural Forest Conservation Program* (in Chinese) (China's Forestry Publishing House, Beijing, 1999).
34. Xinhua News Agency http://english.people.com.cn/200403/31/eng20040331_139043.shtml (2004).
35. WWF-China <http://www.wwfchina.org/english/print.php?loca=159> (2003).

Supplementary information accompanies the paper on www.nature.com/nature.

Acknowledgements We would like to thank Kim Borland, Mason Bradbury, Xiaodong Chen, Jayson Eageler, Linda Fortin, Guangming He, Edward Laurent, Shuxin Li, Wei Liu, Zhiyun Ouyang, Nathan Pfost, William Taylor, Brent Wheat and Daniel Wierich for their helpful assistance, and Rory Howlett and three anonymous reviewers for constructive suggestions. We gratefully acknowledge financial support from the National Science Foundation (Biocomplexity in the Environment), the National Institute of Child Health and Human Development (R01 HD39789), the Michigan Agricultural Experiment Station, and the National Natural Science Foundation of China (30428028).

PROGRESS

Strong present-day aerosol cooling implies a hot future

Meinrat O. Andreae¹, Chris D. Jones² & Peter M. Cox³

Atmospheric aerosols counteract the warming effects of anthropogenic greenhouse gases by an uncertain, but potentially large, amount. This in turn leads to large uncertainties in the sensitivity of climate to human perturbations, and therefore also in carbon cycle feedbacks and projections of climate change. In the future, aerosol cooling is expected to decline relative to greenhouse gas forcing, because of the aerosols' much shorter lifetime and the pursuit of a cleaner atmosphere. Strong aerosol cooling in the past and present would then imply that future global warming may proceed at or even above the upper extreme of the range projected by the Intergovernmental Panel on Climate Change.

'Climate sensitivity' measures how strongly the Earth's climate system responds to a given perturbation, and is often expressed as the equilibrium rise in global temperature resulting from a doubling of atmospheric CO₂. Because it is the key parameter that translates scenarios of future atmospheric composition into projections of climate change, accurate estimates of climate sensitivity are essential. Unfortunately, climate models yield a wide range of sensitivities, depending on the parameterizations they contain^{1–3}, and thus cannot reliably constrain the true climate sensitivity. Alternatively, climate sensitivity can be deduced by relating an observed climate change (for example, the global warming over the last century) to an estimated magnitude of forcing. Such forcing estimates are, however, also highly uncertain, mostly because of incomplete understanding of the climate forcing by atmospheric aerosols^{4–6}.

Future changes in the balance of climate forcing factors—such as increasing greenhouse gases (GHG) but decreasing aerosol burdens—mean that historical changes are not sufficient to constrain future projections. The climate will become more dependent on climate sensitivity as the aerosol burden is reduced. Furthermore, the response of the natural carbon cycle to future climate is also dependent on the climate sensitivity, implying large uncertainties in future CO₂ concentrations.

Climatic effects of aerosols

In the first IPCC report⁷, climate change was considered to be driven predominantly by anthropogenic GHG emissions. Aerosol effects on climate were mentioned, but our knowledge was considered inadequate to estimate their magnitude, or even sign. Since then, the number of aerosol-caused climate effects considered and the estimates of their cumulative magnitude have steadily grown.

All aerosol types (sulphates, organics, mineral dust, sea salt, and so on) intercept incoming sunlight, and reduce the energy flux arriving at the Earth's surface, thus producing a cooling⁸. Some aerosols (for example, soot) absorb light and thereby warm the atmosphere, but also cool the surface. This warming of atmospheric layers may also reduce cloudiness, yielding another warming effect. In addition to these 'direct' radiative effects, there are several 'indirect', cloud-mediated effects of aerosols, which all result in cooling: more aerosols produce more, but smaller, droplets in a given cloud, making it more reflective. Smaller droplets are less likely to coalesce into raindrops, and thus the lifetime of clouds is extended, again increasing the

Earth's albedo. Finally, modifications in rainfall generation change the thermodynamic processes in clouds, and consequently the dynamics of the atmospheric 'heat engine' that drives all of weather and climate. The recent tremendous growth in knowledge of the climatic effects of aerosols, along with the emergence of the likelihood of positive feedbacks between climate and the carbon cycle^{9,10}, have transformed the orderly picture of climate change of the early 1990s, dominated by GHG warming, into a complex mix of opposing effects^{11,12}.

Aerosols and climate sensitivity

Constraints on the value of climate sensitivity (expressed as $\Delta T_{2\times\text{CO}_2}$, the equilibrium temperature response to a doubling of CO₂, see Box 1) are sought by two main approaches. The 'bottom-up' approach, used in General Circulation Models (GCMs), relies primarily on improved representation of the feedbacks in the climate system, including ever more complex representations of physical processes within higher-resolution coupled ocean–atmosphere models¹. These efforts have yielded many interesting scientific insights, but have

Box 1 | Climate sensitivity

The problem of predicting global climate change can be symbolically represented by a simple heat balance equation:

$$c \frac{d(\Delta T)}{dt} = \Delta Q - \lambda \Delta T$$

Here ΔT is the global mean temperature change arising from a change in radiative forcing ΔQ . ΔQ represents the total climate forcing (in W m^{-2}) due to changes in natural factors (such as volcanoes and solar variability), as well as human-induced changes in the concentrations of greenhouse gases and aerosols. The long-term equilibrium response to a radiative forcing (such as doubling of CO₂) is given by the parameter λ as follows: $\Delta T_{2\times\text{CO}_2} = \Delta Q_{2\times\text{CO}_2} / \lambda$, where $\Delta Q_{2\times\text{CO}_2} = 3.7 \text{ W m}^{-2}$. λ itself depends on many climate feedback processes, such as those arising from changes in water vapour, snow cover and clouds. The left-hand side of this equation is a heat storage term which determines how quickly the climate system approaches this equilibrium state. The heat capacity c can be estimated from observations of ocean heat uptake²⁴ and recent warming trends¹³ as $1.1 \pm 0.5 \text{ GJ m}^{-2} \text{ K}^{-1}$. However, there is a wide range in projections of future climate change primarily because of uncertainties in both λ and the future ΔQ .

¹Max Planck Institute for Chemistry, PO Box 3060, Mainz 55131, Germany. ²Hadley Centre for Climate Prediction and Research, Met Office, Fitzroy Road, Exeter EX1 3PB, UK.

³Centre for Ecology and Hydrology, Winfrith, Dorset DT2 8ZD, UK.

failed to reduce the uncertainty in climate sensitivity. The first IPCC report⁷ quoted a range of $\Delta T_{2\times\text{CO}_2} = 1.5\text{--}4.5\text{ K}$, which remains essentially unchanged in the Third Assessment Report of the IPCC¹ (IPCC-TAR). Some groups are now attempting to determine the relative likelihood of different climate sensitivities using the accuracy of simulations of current climate as weighting factors for an ensemble of climate projections². However, GCM parameters are sufficiently uncertain that a recent 'grand ensemble' of more than 2,000 climate change experiments³—all using the same GCM—has yielded sensitivities ranging from below 2 K to more than 11 K.

An alternative 'observationally based' approach⁴ makes use of the observed global warming of $\sim 0.7\text{ K}$ over the 20th century, and of 0.4 K from 1940 to 2000 (ref. 13). The equation shown in Box 1 then offers a means to estimate climate sensitivity, given the net radiative forcing ΔQ over the same period:

$$\Delta T_{2\times\text{CO}_2} = 3.7 \frac{\Delta T}{\Delta Q - c \frac{d(\Delta T)}{dt}} \quad (1)$$

ΔQ is the sum of the relatively well-known GHG forcing ($+2.4 \pm 0.3\text{ W m}^{-2}$ from 1750 to 2000; ref. 1), and the very poorly quantified, but potentially substantial, cooling from anthropogenic aerosols. Consequently, equation (1) shows us that a larger aerosol cooling over the historical period (and thus a smaller net forcing) implies a more sensitive climate (Fig. 1).

Climate 'protection' and future uncertainty

The range of aerosol forcings predicted by 'forward' models, using our best knowledge on the atmospheric aerosol burden and its climate effects, is vast¹⁴, from 0 to -4.4 W m^{-2} . Thus, even if we ignore the implied possibility of a net cooling forcing over the past century, we find that adding the aerosols' effects to those of the GHG yields a net forcing that extends from the full GHG forcing down to a zero net forcing over the last century.

Although the models may disagree about the magnitude of the aerosol effect, they all agree that the net effect is cooling, and that aerosols have therefore 'protected' us from some of the greenhouse warming. The price for this 'climate protection' is, however, great uncertainty about the true magnitude of the climate change we can expect in the future. If the mix of future forcings remains the same as in the past, precise knowledge of λ would not be necessary: historical changes would constrain the future¹⁵. However, because we expect the proportions of GHG and aerosols to change in the future, past changes become a much weaker constraint on future behaviour (Box 2).

The twentyfirst-century climate will therefore suffer the treble

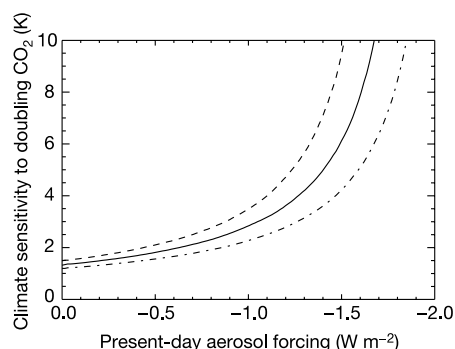


Figure 1 | Climate sensitivity required to explain the observed 1940–2000 warming as a function of the strength of aerosol radiative cooling. The solid line represents results using the central estimate of heat capacity ($1.1 \pm 0.5\text{ GJ m}^{-2}\text{ K}^{-1}$) from Levitus *et al.*²⁴, and the dashed (dot-dashed) lines represent the higher (lower) limit of this heat capacity. More details of the model are given in Box 3.

hit of an increasing warming from greenhouse gases, a decreasing cooling from aerosols, and positive feedbacks from the carbon cycle, whereby increased temperatures cause accelerated release of soil carbon by decomposition⁹. The effects of anthropogenic aerosols have created great uncertainty in our knowledge of the climate sensitivity to increasing greenhouse gases. Do we live in a world with weak aerosol cooling and thus low climate sensitivity, in which case future climate change may be expected to be relatively benign? Or do we live in a highly forced, highly sensitive world with a very uncertain and worrying future that may bring a much faster temperature rise than is generally anticipated?

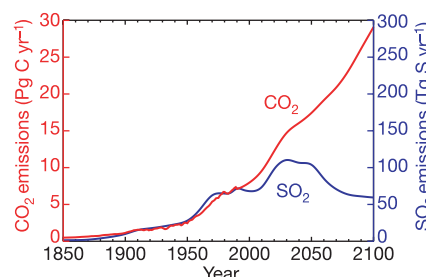
Historical constraints

To explore this issue more quantitatively, we use a deliberately simplistic approach to illustrate the impact of the uncertainties on projections of future climate (see Box 3). We apply two observational constraints: the model should reproduce both the observed global warming and the CO_2 increase from 1940–2000. The observed warming constraint yields a relationship between aerosol cooling and climate sensitivity (Fig. 1), indicating that $\Delta T_{2\times\text{CO}_2}$ is just 1.3 K for zero aerosol forcing, but exceeds 10 K for $\Delta Q_{\text{aeros}} = -1.7\text{ W m}^{-2}$ (ΔQ_{aeros} is the sum of all aerosol forcings). A climate sensitivity of 10 K is large, but cannot be ruled out by observations. Climate sensitivity diagnosed by this model rapidly becomes unphysically large and then negative as the aerosol forcing exceeds -1.7 W m^{-2} . We note that recent studies tend to estimate the sum of aerosol forcings to be in the range -1 to -2 W m^{-2} , that is, in the region of sharply increasing and highly uncertain $\Delta T_{2\times\text{CO}_2}$ (refs 11, 12, 16, 17).

In a similar vein, the historical CO_2 rise sets a joint constraint on the parameter determining the CO_2 -fertilization of photosynthesis ($C_{0.5}$, the half-saturation concentration for photosynthesis), and the parameter determining the sensitivity of soil respiration to temperature (q_{10} —the factor by which decomposition accelerates for each

Box 2 | Future aerosol scenarios

The SRES emissions scenarios²⁵ used in the IPCC-TAR all suggest that aerosol emission by the middle of this century will be near or below present levels. Because aerosols are very short-lived in the atmosphere—lifetimes of days compared with decades for the greenhouse gases—they do not accumulate and the burden is almost proportional to the emissions. Consequently, as we clean up our vehicles and smokestacks to provide cleaner air and improve air quality, the aerosol loading of the atmosphere will decrease. Even population growth and increasing industrialization in the developing countries will do little to change this outcome. We are already in the process of revising downward our projections of aerosol emissions from China and other developing countries, as they are introducing cleaner technology faster than had been anticipated a decade or so ago. Because of the rapidly growing knowledge of the very serious health effects of aerosols²⁸ we expect that regulatory efforts will act to reduce aerosol emissions even more rapidly than anticipated when the SRES scenarios were developed.



Box 2 Figure | Historical CO_2 and SO_2 emissions from 1850–2000, followed by projected values to the year 2100 from the SRES²⁵ A2 scenario.

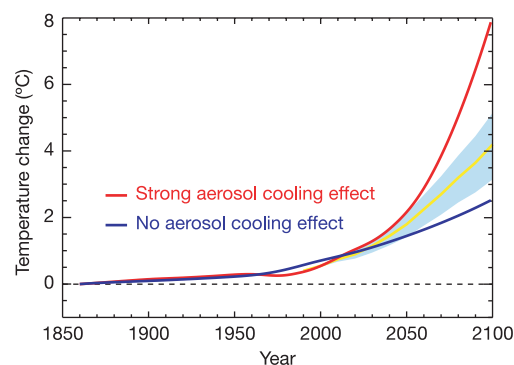


Figure 2 | Temperature change simulated by the simple model for the period 1850 to 2100. Two extreme cases are shown: strong present-day aerosol cooling consistent with 'forward' studies of aerosol effects on climate but with a climate sensitivity not ruled out by observations (red line, $Q_{\text{aeros}} = -1.7 \text{ W m}^{-2}$), and the case of no aerosol cooling effect (blue line). The shading and the yellow line represent the range and central projection given in IPCC-TAR, based on the same scenario used in these calculations (scenario A2, from ref. 25).

10-K warming^{18,19}). Both high fertilization + high q_{10} and low fertilization + low q_{10} are consistent with the historical rise of CO_2 and temperature, but imply different responses of the land carbon cycle to future climates and therefore very different magnitudes of the carbon cycle feedback.

Future projections

We have run the simple model on to 2100 for a range of scenarios from the IPCC's Special Report on Emissions Scenarios (SRES, see Box 2) (Figs 2 and 3). We find that a large uncertainty range of temperature increase is predicted for 2100, and that even by 2050, the model runs with strong historical aerosol cooling predict a temperature rise from 1850 of as much as 2.2°C .

The implied high climate sensitivities are within the range of sensitivities inferred by recent observational approaches. Analyses of the probability distribution of climate sensitivities that can be deduced from climate observations suggest that there is a significant probability that the true climate sensitivity is in excess of 4 K (refs 5, 6), and maybe as high as 10 K. Recent analyses of the palaeoclimatic record also suggest fairly high climate sensitivity^{20,21}.

When we include the uncertainty caused by the choice of emission scenarios, we find that the range considered most plausible in IPCC-TAR ($2.3\text{--}4.9^\circ\text{C}$ from 1850–2100) can be obtained only for aerosol forcings considerably weaker than predicted by current forward models (Fig. 3), which tend to estimate the sum of aerosol forcings to be in the range -1 to -2 W m^{-2} (refs 11, 12, 16, 17). Ominously, Fig. 3a shows temperature increases in excess of 6°C for the climate sensitivity implied by the central estimate of aerosol forcing (-1.5 W m^{-2}), and for all but the most optimistic emission scenario. Such an enormous increase would be comparable to the temperature change from the previous ice age to the present. Furthermore, the overall uncertainty is dominated by climate sensitivity and hence historical aerosol forcing: Fig. 3a shows that the warming range for a given scenario (for example, $2.5\text{--}7.9 \text{ K}$ for scenario A2) is greater than the range across scenarios for a given climate sensitivity ($6.8\text{--}9.6 \text{ K}$ at its widest).

Part of the reason for this extraordinary sensitivity of future projections to the historical aerosol forcing is due to the impact of the carbon cycle feedback on projected CO_2 levels (Fig. 3b). The extent to which the land carbon cycle amplifies future CO_2 increase depends critically on climate sensitivity (Fig. 4). The positive climate–carbon cycle feedback increases markedly with climate

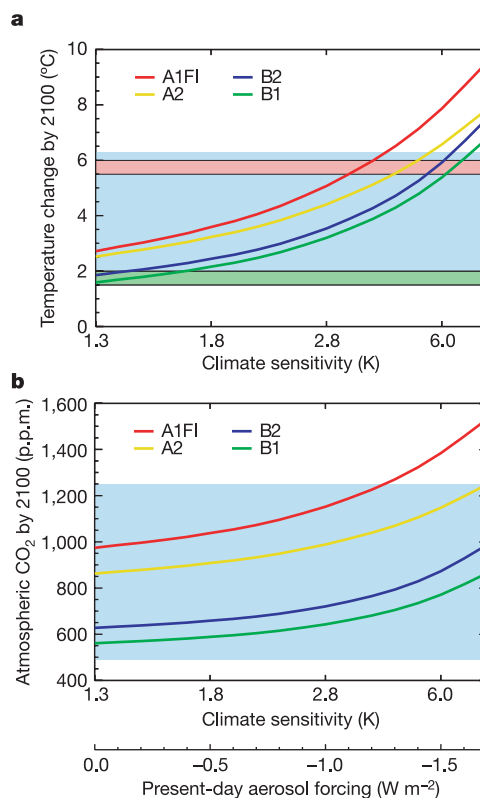


Figure 3 | Modelled temperature change and CO_2 increase by 2100 under different development scenarios. **a**, Temperature rise by the year 2100 for the various SRES scenarios²⁵ as a function of present-day aerosol cooling. The horizontal green bar indicates the threshold of 'dangerous anthropogenic interferences' in the sense of the United Nations Framework Convention on Climate Change, using an estimate of $1.5\text{--}2^\circ\text{C}$ for this value, based on arguments by different groups^{26,27}. At all but the very lowest climate sensitivities this level will be exceeded unless GHG emissions are reduced below those in the SRES B1 and B2 scenarios (from ref. 25). The pink bar indicates the temperature change between ice ages and interglacials²⁰. **b**, Same as **a** but for atmospheric CO_2 by 2100. The shaded areas represent the IPCC-TAR range across models and scenarios.

Box 3 | Simple climate–carbon cycle model

We use a zero-dimensional climate–carbon cycle model²⁹, which updates the global temperature using the equation in Box 1 and accounts for potentially large positive carbon cycle feedbacks by updating CO_2 interactively on the basis of the emissions scenario. It uses a simple fit to the ocean and land uptake of CO_2 derived from the Hadley Centre's climate–carbon cycle GCM⁹, but with alternative sets of possible land sensitivity parameters chosen to fit the observed CO_2 rise. The land carbon cycle responses produced by this simple fit therefore span the range simulated by other potentially realistic models. The size of the climate–carbon cycle feedback depends critically on the opposing effects of CO_2 -fertilization on plant growth, and enhanced soil decomposition as the climate warms. The latter is dependent on the degree of climate warming, as well as the sensitivity of soil respiration to temperature^{18,29}.

The major anthropogenic forcings are considered: CO_2 , other well-mixed GHGs, and sulphate aerosols. The radiative forcing from CO_2 and other well-mixed GHGs are derived from well-known formulae¹. The radiative forcing from sulphate aerosols is assumed to be proportional to global mean sulphate loading, which in turn is assumed to be proportional to SO_2 emissions. To avoid undue influence of other forcing factors (in particular natural forcing from solar and volcanic sources) we consider just the portion of the historical record that is dominated by anthropogenic influence—namely 1940 to present³⁰.

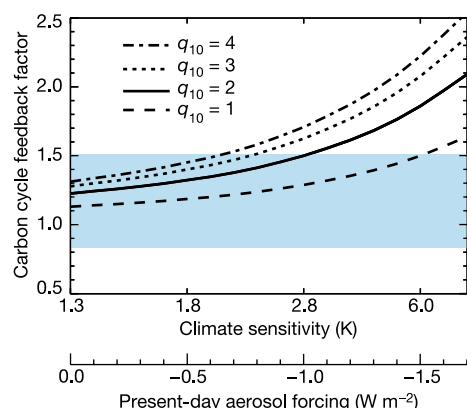


Figure 4 | Strength of climate-carbon cycle feedback as a function of climate sensitivity. The feedback factor is defined as the CO₂ concentration rise projected between 1860 and 2100, divided by the CO₂ rise predicted in the absence of climate effects on the carbon cycle. Results are shown for the A2 scenario as a function of climate sensitivity/present-day aerosol cooling, for various sensitivities of soil respiration to temperature. The shaded area represents the range of feedback factor estimated from the IPCC-TAR range of CO₂ concentrations relative to the standard A2 concentration scenario. q_{10} values of 1, 2, 3 and 4 correspond to $C_{0.5}$ values of 295, 485, 676, 866 p.p.m., respectively, in order for the model to recreate the observed CO₂ record.

sensitivity, especially if soil decomposition is more sensitive to temperature (that is, for higher q_{10} values). With a best estimate²² of $q_{10} = 2$, we find that for climate sensitivities greater than 3 °C, the carbon cycle feedback will accelerate CO₂ growth by more than 50%. This dependence of the carbon cycle feedback strength on climate sensitivity may explain a large part of the divergence amongst the first generation climate-carbon cycle GCMs²³.

Thus research over the past decade has shown evidence of the importance of a considerable number of aerosol climatic effects, which on balance cool the Earth and have therefore reduced the effect of greenhouse warming. Because of the stabilizing emission of aerosols and their short lifetime, this 'protection' will diminish in the future, leaving us vulnerable to both greater climate change and greater uncertainty. Incomplete consideration of aerosols in current climate models may have led to underestimation of the true climate sensitivity. We cannot quantitatively assess the probability of a given climate sensitivity within the limited scope of this paper, but our analysis suggests that there is a possibility that climate change in the twenty-first century will follow the upper extremes of current IPCC estimates, and may even exceed them. Such a degree of climate change is so far outside the range covered by our experience and scientific understanding that we cannot with any confidence predict the consequences for the Earth system.

To reduce these uncertainties a multi-pronged approach is needed. First, there is a great need for *in situ* studies that investigate the response of cloud microphysics and dynamics to enhanced aerosol concentrations. Second, at the regional and global scale, the effects of aerosols on cloud properties and abundance must be studied using remote-sensing data from the newly available and upcoming satellite sensors. Third, parameterizations of cloud processes and feedbacks in GCMs must be improved. Finally, uncertainties in feedbacks that are strongly dependent on climate sensitivity, such as the carbon cycle feedback, must also be reduced, through process studies and model improvements.

- Houghton, J. T. *et al.* in *Climate Change 2001: The Scientific Basis. Contribution of Working Group I to the Third Assessment Report of the Intergovernmental Panel on Climate Change* (Cambridge Univ. Press, UK/New York, NY, 2001).
- Murphy, J. M. *et al.* Quantification of modelling uncertainties in a large ensemble of climate change simulations. *Nature* **430**, 768–772 (2004).
- Stainforth, D. A. *et al.* Uncertainty in predictions of the climate response to rising levels of greenhouse gases. *Nature* **433**, 403–406 (2005).
- Gregory, J. M., Stouffer, R. J., Raper, S. C. B., Stott, P. A. & Rayner, N. A. An observationally based estimate of the climate sensitivity. *J. Clim.* **15**, 3117–3121 (2002).
- Forest, C. E., Stone, P. H., Sokolov, A. P., Allen, M. R. & Webster, M. D. Quantifying uncertainties in climate system properties with the use of recent climate observations. *Science* **295**, 113–117 (2002).
- Knutti, R., Stocker, T. F., Joos, F. & Plattner, G. K. Constraints on radiative forcing and future climate change from observations and climate model ensembles. *Nature* **416**, 719–723 (2002).
- Houghton, J. T., Jenkins, G. J. & Ephraums, J. J. *Climate Change: The IPCC Assessment* (Cambridge Univ. Press, Cambridge, UK, 1990).
- Charlson, R. J. *et al.* Climate forcing by anthropogenic aerosols. *Science* **255**, 423–430 (1992).
- Cox, P. M., Betts, R. A., Jones, C. D., Spall, S. A. & Totterdell, I. J. Acceleration of global warming due to carbon-cycle feedbacks in a coupled climate model. *Nature* **408**, 184–187 (2000).
- Friedlingstein, P. *et al.* Positive feedback between future climate change and the carbon cycle. *Geophys. Res. Lett.* **28**, 1543–1546 (2001).
- Lohmann, U. & Feichter, J. Global indirect aerosol effects: a review. *Atmos. Chem. Phys.* **5**, 715–737 (2005).
- Hansen, J. *et al.* Climate forcings in Goddard Institute for Space Studies SI2000 simulations. *J. Geophys. Res.* **107**, 4347, doi:10.1029/2001JD001143 (2002).
- Folland, C. K. *et al.* Global temperature change and its uncertainties since 1861. *Geophys. Res. Lett.* **28**, 2621–2624 (2001).
- Anderson, T. L. *et al.* Climate forcing by aerosols—a hazy picture. *Science* **300**, 1103–1104 (2003).
- Allen, M. R., Stott, P. A., Mitchell, J. F. B., Schnur, R. & Delworth, T. L. Quantifying the uncertainty in forecasts of anthropogenic climate change. *Nature* **407**, 617–620 (2000).
- Koch, D. Transport and direct radiative forcing of carbonaceous and sulfate aerosols in the GISS GCM. *J. Geophys. Res.* **106**, 20311–20332 (2001).
- Haywood, J. & Boucher, O. Estimates of the direct and indirect radiative forcing due to tropospheric aerosols: A review. *Rev. Geophys.* **38**, 513–543 (2000).
- Knorr, W., Prentice, I. C., House, J. I. & Holland, E. A. Long-term sensitivity of soil carbon turnover to warming. *Nature* **433**, 298–301 (2005).
- Powlson, D. Will soil amplify climate change? *Nature* **433**, 204–205 (2005).
- Alley, R. B. Palaeoclimatic insights into future climate challenges. *Phil. Trans. R. Soc. Lond. A* **361**, 1831–1848 (2003).
- Jenkyns, H. C., Forster, A., Schouten, S. & Damste, J. S. S. High temperatures in the Late Cretaceous Arctic Ocean. *Nature* **432**, 888–892 (2004).
- Jones, C. D. & Cox, P. M. Constraints on the temperature sensitivity of global soil respiration from the observed interannual variability in atmospheric CO₂. *Atmos. Sci. Lett.* **2**, doi:10.1006/asle.2001.0041 (2001).
- Friedlingstein, P., Dufresne, J. L., Cox, P. M. & Rayner, P. How positive is the feedback between climate change and the carbon cycle? *Tellus B* **55**, 692–700 (2003).
- Levitus, S., Antonov, J. I., Boyer, T. P. & Stephens, C. Warming of the world ocean. *Science* **287**, 2225–2229 (2000).
- Nakicenovic, N. & Swart, R. (eds) *Special Report on Emissions Scenarios* (Cambridge Univ. Press, Cambridge, UK, 2000).
- O'Neill, B. C. & Oppenheimer, M. Dangerous climate impacts and the Kyoto protocol. *Science* **296**, 1971–1972 (2002).
- Hansen, J. E. A slippery slope: How much global warming constitutes "Dangerous anthropogenic interference"? *Clim. Change* **68**, 269–279 (2005).
- Pope, C. A. *et al.* Lung cancer, cardiopulmonary mortality, and long-term exposure to fine particulate air pollution. *J. Am. Med. Assoc.* **287**, 1132–1141 (2002).
- Jones, C. D., Cox, P. & Huntingford, C. Uncertainty in climate-carbon-cycle projections associated with the sensitivity of soil respiration to temperature. *Tellus B* **55**, 642–648 (2003).
- Stott, P. A. *et al.* Attribution of twentieth century temperature change to natural and anthropogenic causes. *Clim. Dyn.* **17**, 1–21 (2001).

Acknowledgements C.D.J. was supported by the UK DEFRA Climate Prediction Program.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to M.O.A. (andreae@mpch-mainz.mpg.de).

Short-lived orogenic cycles and the eclogitization of cold crust by spasmodic hot fluids

Alfredo Camacho¹, James K. W. Lee¹, Bastiaan J. Hensen² & Jean Braun³

Collision tectonics and the associated transformation of continental crust to high-pressure rocks (eclogites) are generally well-understood processes, but important contradictions remain between tectonothermal models and petrological-isotopic data obtained from such rocks. Here we use ^{40}Ar – ^{39}Ar data coupled with a thermal model to constrain the time-integrated duration of an orogenic cycle (the burial and exhumation of a particular segment of the crust) to be less than 13 Myr. We also determine the total duration of associated metamorphic events to be ~ 20 kyr, and of individual heat pulses experienced by the rocks to be as short as 10 years. Such short timescales are indicative of rapid tectonic processes associated with catastrophic deformation events (earthquakes). Such events triggered transient heat advection by hot fluid along deformation (shear) zones, which cut relatively cool and dry subducted crust. In contrast to current thermal models that assume thermal equilibrium and invoke high ambient temperatures in the thickened crust, our non-steady-state cold-crust model satisfactorily explains several otherwise contradictory geological observations.

Transformation of crustal rocks to eclogite takes place during burial by the effect of deformation and/or fluid infiltration^{1–4}. However, many exposed high-pressure terranes are only partly transformed to eclogite, suggesting that buoyancy may be a controlling factor in their rate of exhumation^{1,5–8}. The Southern Caledonides of Norway, including the Lindås nappe in the Bergen arcs, are one of the best-documented examples of a continental-collision zone containing high- and ultrahigh-pressure rocks. Eclogites, confined to Caledonian (~ 425 Myr ago)⁹ shear zones within Proterozoic unreacted dry granulites and anorthosites, formed at temperatures of $\sim 700^\circ\text{C}$ and depths of ~ 60 km (refs 1, 10–12). Phlogopite in undeformed ultramafic lenses surrounded by Caledonian eclogite, however, preserves Proterozoic Rb–Sr ages (~ 850 Myr)¹³. Because the temperature of eclogite formation was assumed to reflect the ambient thermal regime^{13,14} in the subducted crust at 60 km depth, this unexpected result was inferred to indicate a Rb–Sr phlogopite closure temperature of $\sim 700^\circ\text{C}$ under dry, static conditions¹³. Using the ^{40}Ar – ^{39}Ar dating technique, we explore the alternative possibility that the country rocks did not reach the high temperatures recorded by the eclogites, and investigate the resulting geological implications.

The ^{40}Ar – ^{39}Ar method is especially well-suited to thermochronological studies of the Earth's crust, and allows us to uniquely constrain the thermal history of the untransformed, granulite-facies country rocks in the Bergen arcs. Various minerals (see Supplementary Tables 1–6 for all analyses) were separated from two ultramafic lenses surrounded by eclogitic shear zones on Holsnøy island (Fig. 1); these were the same lenses examined in the Rb–Sr study¹³. The Alvjellet lens (samples Alv6 and Alv7) is a 25×30 m outcrop, whereas the Hundskjeften lens (sample Hunds14) is 10×20 m. Phlogopite in these lenses is in textural equilibrium with other minerals and is compositionally uniform from core to rim, suggesting negligible compositional re-equilibration after granulite-

facies metamorphism at ~ 930 Myr ago^{13,15}. Amphiboles from Alv6 are also pristine, uniform in composition, and do not contain either fluid or mineral inclusions, as demonstrated by optical microscopy, backscattered electron imaging, X-ray mapping and electron microprobe analyses (Supplementary Table 1). The Hundskjeften lens also preserves its original magmatic assemblage, but locally is partially overprinted by Caledonian garnet and clinopyroxene forming a rind a few centimetres thick along the margins.

Argon dating and isotope systematics

^{40}Ar – ^{39}Ar laser step-heating experiments on two amphibole aliquots of different grain size from sample Alv6 yielded age spectra (Fig. 2) partially affected by argon (^{40}Ar) that did not form from radioactive decay of ^{40}K within the crystal (here termed 'excess'¹⁶ $^{40}\text{Ar}_\text{E}$). Ca/K and Cl/K ratios for 99.99% of the ^{39}Ar released are concordant, indicating that argon is derived exclusively from amphibole with uniform chemical and isotopic composition. The larger grain (~ 350 μm diameter) is less affected by $^{40}\text{Ar}_\text{E}$ than the smaller grain (~ 100 μm diameter), consistent with the uptake of excess argon as predicted by diffusion theory. The plateau segments (representing $>70\%$ of ^{39}Ar released) for both aliquots yield identical ages of ~ 885 Myr, which are slightly younger than the U–Pb zircon age of ~ 930 Myr for regional granulite-facies metamorphism¹⁵. Total-fusion ^{40}Ar – ^{39}Ar ages of eight amphibole crystals range from ~ 885 to 915 Myr, also reflecting varying degrees of $^{40}\text{Ar}_\text{E}$. Phlogopite from the two lenses also yielded ^{40}Ar – ^{39}Ar age spectra partially affected by $^{40}\text{Ar}_\text{E}$ (Fig. 3); dates monotonically decrease from ~ 975 to 900 Myr ago in the lower-temperature steps to plateau-like segments in each age spectrum. Together, the plateau-like segments from all three samples span a range of ages from ~ 820 to 895 Myr, similar to the Rb–Sr phlogopite ages of ~ 840 – 860 Myr (ref. 13). In addition, the integrated ^{40}Ar – ^{39}Ar ages from 18 phlogopite crystals (1,600–200 μm

¹Geological Sciences and Geological Engineering, Queen's University, Kingston, Ontario K7L 3N6, Canada. ²School of Biology, Earth and Environmental Sciences, University of New South Wales, Sydney, New South Wales 2052, Australia. ³Géosciences Rennes, CNRS UMR 6118, Université de Rennes 1, Rennes F-35042, France.

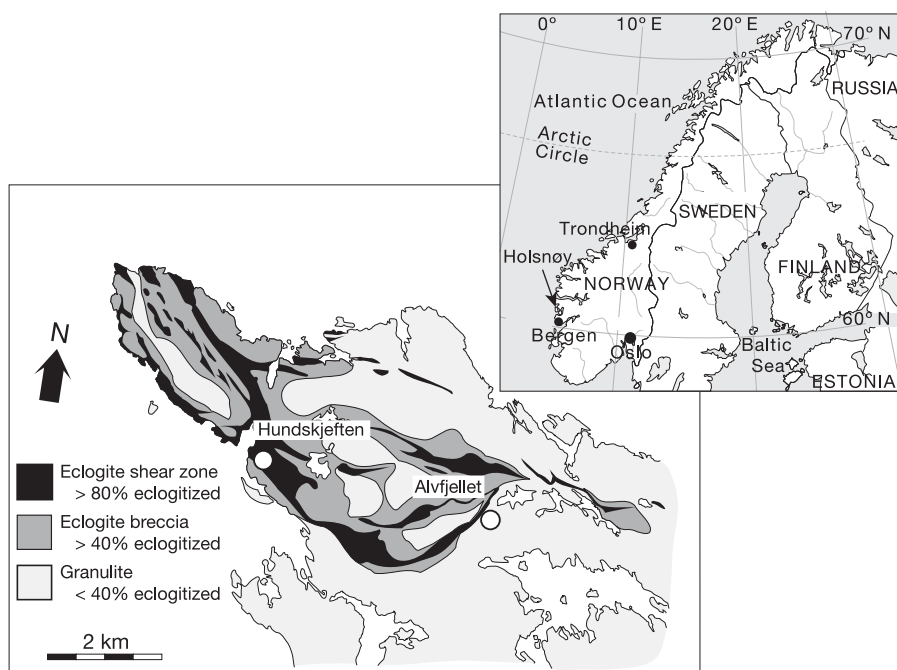


Figure 1 | Geological map of Holsnøy island, northwest of Bergen, western Norway. This shows the distribution of the Caledonian (~425 Myr ago) eclogite-facies overprint and the Alvjellet and Hundskjeften localities within the eclogite-facies rocks. After refs 40 and 41.

diameter) show that the ages of the larger grains are less affected by $^{40}\text{Ar}_\text{E}$ than those of the smaller grains (Fig. 4), also consistent with the uptake of excess argon by diffusion. As well as the step-heating experiments, narrow trenches (~30 μm wide) parallel to the grain margin were ablated in three phlogopite grains, using an ultraviolet laser¹⁷. The age profiles from the three grains (Fig. 5) show that the ages decrease inwards away from the grain boundary, confirming the intragrain age distributions inferred from the step-heating experiments. Our data clearly demonstrate that amphibole and phlogopite are affected by $^{40}\text{Ar}_\text{E}$ and must have been open to argon diffusion since the Proterozoic.

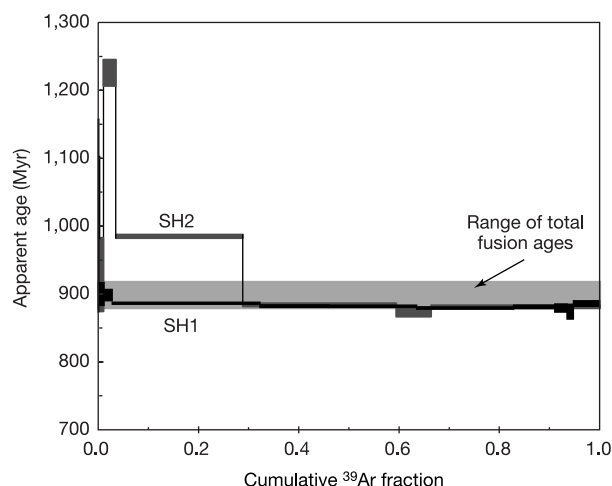


Figure 2 | ^{40}Ar – ^{39}Ar release spectra for two step-heated amphibole aliquots of different grain size from the Alvjellet lens (sample Alv6) on Holsnøy island. Both age spectra are affected by excess argon ($^{40}\text{Ar}_\text{E}$) in the lower-temperature steps to varying degrees. Aliquot SH1 (~350 μm diameter) and SH2 (~100 μm diameter) yield identical plateau ages of 885 ± 2 Myr (2σ ; mean square of the weighted deviates MSWD = 1.3; 97.2% of ^{39}Ar) and 882 ± 3 Myr (2σ ; MSWD = 0.7; 71.3% of ^{39}Ar), respectively, and constrain the cooling age for granulite-facies metamorphism. The shaded band represents the range of ^{40}Ar – ^{39}Ar total-fusion ages (885–915 Myr) of eight amphibole crystals from the same sample. Box heights are $\pm 1\sigma$.

Because Caledonian metamorphism is the only other recorded thermal event since the Proterozoic¹³, we investigated whether the incorporation of $^{40}\text{Ar}_\text{E}$ in the phlogopites could be linked to that event by analysing the argon concentrations of other minerals in the host rocks (Supplementary Table 6). Caledonian garnet and clinopyroxene along the margin of the Hundskjeften lens contain $^{40}\text{Ar}_\text{E}$ concentrations that are 1–2 orders of magnitude higher than those in the Proterozoic magmatic assemblage in the ultramafic lenses. Thus, we can directly attribute the source of $^{40}\text{Ar}_\text{E}$ to the fluids associated with the Caledonian event, in agreement with previous studies^{18,19}. However, as the untransformed lenses have not been hydrated during the eclogitization event¹³, this suggests that an argon-rich, but water-poor fluid infiltrated along the grain boundaries of the ultramafic

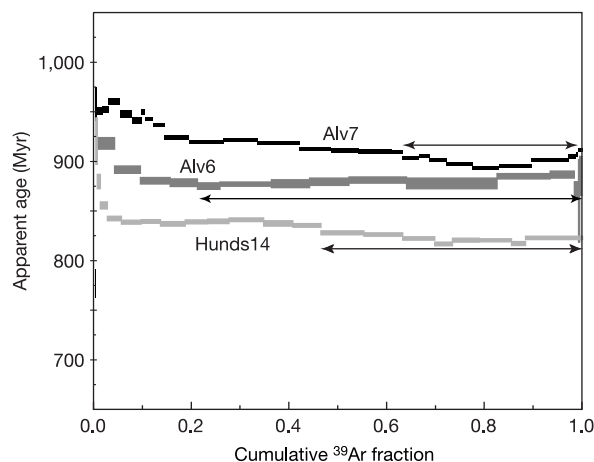


Figure 3 | ^{40}Ar – ^{39}Ar release spectra for step-heated phlogopite from the Alvjellet (Alv6; Alv7) and Hundskjeften (Hunds14) lenses on Holsnøy island. All three age spectra exhibit excess argon ($^{40}\text{Ar}_\text{E}$) in the lower-temperature steps. Alv6 yields a plateau age (spanned by arrow) of 880 ± 4 Myr (2σ ; MSWD = 1.3; 78.71% of ^{39}Ar). Alv7 and Hunds14 do not yield plateaus, but plateau-like segments as indicated by arrows; Alv7 = 897 ± 3 Myr (2σ ; MSWD = 5.6; 37.77% of ^{39}Ar) and Hunds14 = 821 ± 3 Myr (2σ ; MSWD = 2.9; 52.44% of ^{39}Ar). Box heights are $\pm 1\sigma$.

rocks, in contrast with the water-rich fluid in the shear zones. This is, to our knowledge, the first evidence for an open system in the untransformed protolith well beyond the eclogitization front (compare ref. 18).

Modelling uptake of excess ^{40}Ar

The presence of $^{40}\text{Ar}_\text{E}$ in rocks is generally considered to be detrimental to the interpretation of ^{40}Ar – ^{39}Ar age data, but we can take advantage of its presence by examining the distribution of $^{40}\text{Ar}_\text{E}$ in various K-bearing minerals. By considering the effect that $^{40}\text{Ar}_\text{E}$ would have on the ages of coexisting phlogopite and amphibole (such as in Alv6), we can constrain, in a novel way, the temperature and duration of heating of the country rocks during the eclogitization event. The partitioning of argon between different coexisting K-bearing minerals is commonly assumed to be 1. This supposition, however, is questionable when biotite is involved, given its known high argon solubility relative to most other K-bearing minerals²⁰. Although there are no experimental data on the distribution of argon between amphibole and phlogopite, we derived an approximate partition coefficient ($K_d \approx 0.6$), consistent with the above observation, from argon isotopic concentrations of coexisting amphibole and biotite in rocks from central Australia²¹. Raising the age of a phlogopite with 8.1 wt% K from 875 Myr (youngest age peak from the Alv6, Fig. 5) to 1,100 Myr (laser trench at grain margin of Alv6; Fig. 5) requires $5.47 \times 10^{-9} \text{ mol g}^{-1}$ of $^{40}\text{Ar}_\text{E}$. Complete equilibration of the $^{40}\text{Ar}_\text{E}$ into the coexisting amphibole with 1.5 wt% K, using the estimated partition coefficient, increases its apparent age from 885 to 1,521 Myr—much older than any of our measured ^{40}Ar – ^{39}Ar amphibole ages (885–915 Myr). This apparent age discrepancy can be reconciled by considering the partial uptake of $^{40}\text{Ar}_\text{E}$ controlled by diffusion, as discussed below.

The range of experimentally determined amphibole ages (885–915 Myr) in amphibole crystals with a diameter of $\sim 300 \mu\text{m}$ can be produced by the uptake of up to 3.97% of the total $^{40}\text{Ar}_\text{E}$ concentration. Assuming that only the outermost region of an amphibole grain, equivalent to 3.97% of its volume, is affected (which is further supported by the amphibole step-heating spectrum SH1 in Fig. 2), this translates into a mean radial diffusion distance ($\bar{x}$) for $^{40}\text{Ar}_\text{E}$ diffusion of $2.01 \mu\text{m}$ for a $\sim 300 \mu\text{m}$ grain (assuming a spherical diffusion geometry). In the largest phlogopite grains, only the first 20% of the age spectra is affected by $^{40}\text{Ar}_\text{E}$ (Fig. 3), which volumetrically corresponds to $\bar{x} = 79.2 \mu\text{m}$ (cylindrical geometry) for a

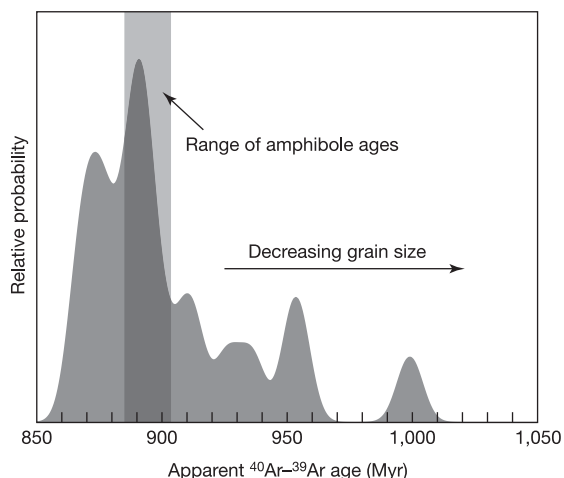


Figure 4 | Cumulative probability diagram of integrated ^{40}Ar – ^{39}Ar ages for phlogopite from the Alv6 lens. The figure shows the correlation between older ages and decreasing grain size. Grains range from 1,600 to $200 \mu\text{m}$ in diameter ($n = 18$). The shaded band represents the same range of amphibole ages as shown in Fig. 2.

$1,500 \mu\text{m}$ diameter grain. The $\bar{x}$ of Ar (for either phlogopite or amphibole) can be used in conjunction with the appropriate diffusion parameters (see legends to Figs 5 and 6) to generate a temperature (T) versus time (t) curve composed of an infinite number of (T – t) pairs corresponding to this $\bar{x}$. By plotting the T – t curves for both minerals, we obtain the conditions under which both isotopic data sets are internally consistent at the intersection of the two curves, where $t = 18 \text{ kyr}$ and $T = 526^\circ\text{C}$ (Fig. 6). This intersection represents an integrated thermal spike experienced by the interiors of the ultramafic lenses due to heat conducted from the hot fluids in the surrounding eclogitized shear zones.

The temperature spike is consistent with the preservation of the

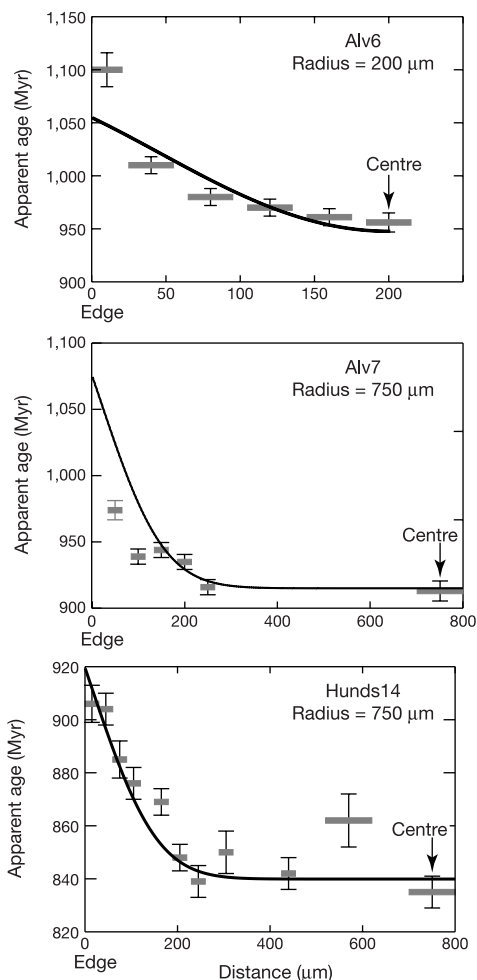


Figure 5 | Apparent ^{40}Ar – ^{39}Ar age versus distance profiles across phlogopite from Alv6, Alv7 and Hunds14. On the horizontal axis, the length of the symbol represents the width of the trench analysed; the error on the age is $\pm 2\sigma$. Superimposed on these profiles are the theoretically derived apparent-age curves for volume diffusion of $^{40}\text{Ar}_\text{E}$ into different-sized phlogopites from the grain boundary due to a thermal spike of 526°C lasting 18 kyr. Observed maximum diffusion distances of $200 \mu\text{m}$ are entirely consistent with the mean (integrated) diffusion distances of $79 \mu\text{m}$ discussed in the text. $^{40}\text{Ar}_\text{E}$ concentrations in the grain-boundary network are: Hunds14, $1.00 \times 10^{-8} \text{ mol g}^{-1}$; Alv6, $1.37 \times 10^{-8} \text{ mol g}^{-1}$; Alv7, $1.25 \times 10^{-8} \text{ mol g}^{-1}$. These concentrations are remarkably similar to each other, suggesting that the fluids served as a common argon reservoir. Argon diffusion parameters for phlogopite¹² were used (activation energy $E = 242.3 \text{ kJ mol}^{-1}$ and pre-exponential coefficient $D_0 = 7.5 \times 10^{-5} \text{ m}^2 \text{ s}^{-1}$), together with a cylindrical diffusion geometry. Note that some of the ages within the profile are anomalously high relative to their position in the grain, and probably reflect the introduction of $^{40}\text{Ar}_\text{E}$ along structural defects²⁹.

diffusion profiles in phlogopite (Fig. 5). To preserve these age profiles, the ambient temperature of the rocks must have been well below 700 °C, because at that temperature the profiles would homogenize within 30–50 yr. Our new temperature estimate, however, indicates that the high temperature (~700 °C) determined for eclogite minerals within the shear zones cannot apply to the country rocks, and therefore, the common assumption that the temperature derived from metamorphic assemblages in shear zones reflects the ambient thermal regime may not be correct. Our calculated integrated heating time of only 18 kyr is much shorter than previous estimates for the life of the shear zones based on diffusion profiles in garnet (1–2 Myr)^{12,22}, although this difference may not be significant, given the uncertainties. However, the calculated duration of heating is consistent with the ~1 kyr period inferred from models of fracture-controlled eclogitization⁴. Our results not only are compatible with current experimental data for argon diffusion in phlogopite and amphibole, but also provide compelling evidence for diffusion-controlled argon gain under dry, static conditions (as the lenses are competent units within zones of deformation). Hence, it is not necessary to invoke anomalously high closure temperatures or the inhibition of intragranular diffusion under dry conditions¹³ to explain the preservation of protolith ages.

Tectonic constraints

To constrain the duration of the complete orogenic cycle (subduction and exhumation of a particular segment of the crust), we combine a thermal model of subduction and exhumation with the diffusion modelling of the argon isotopes. Before discussing the modelling in detail, we first briefly review the tectonic regime during the late Caledonian.

Closure of the Iapetus Ocean during oblique collision between Baltica and Laurentia in Mid Silurian to Early Devonian times (the Scandian event of the Caledonian orogeny) led to the high-pressure

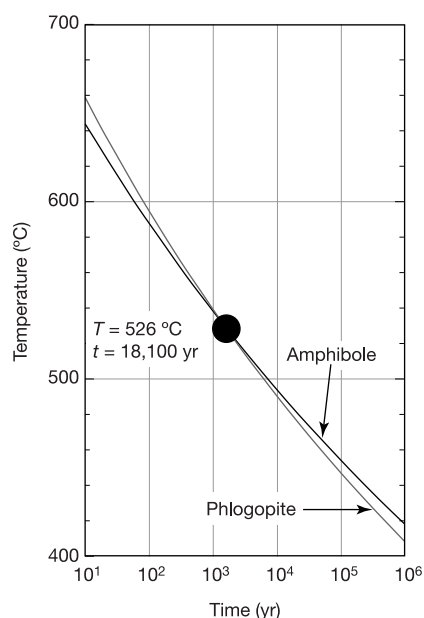


Figure 6 | Relationship between time and temperature required for a 300- μm -diameter amphibole grain and a 1,500- μm -diameter phlogopite grain to incorporate 3.97 vol.% and 20 vol.% $^{40}\text{Ar}_\text{E}$, respectively. The calculated curves incorporate a K_d of 0.6 and represent the corresponding conditions that yield respective mean diffusion distances of 2.01 μm (amphibole) and 79 μm (phlogopite). The two data sets are internally consistent only at the intersection of the two curves (time $t = 18$ kyr and temperature $T = 526$ °C). Argon diffusion parameters for phlogopite (as stated in Fig. 5) and amphibole⁴³ ($E = 268.3$ kJ mol⁻¹ and $D_0 = 2.4 \times 10^{-6}$ m² s⁻¹) were used.

subduction of the Baltic margin^{23,24}. Before this collision, crustal movement is considered to be predominantly subhorizontal; tectonic events that transported crust to depth were confined to the continent–ocean transition zone of the Baltoscandian margin (Late Cambrian) and to the eastern margin of Laurentia (Mid to Late Ordovician)^{24,25}. In the Bergen arcs, the Lindås nappe is considered to have a probable Baltica ancestry, and burial from depths of <30 km to ~60 km has therefore been linked with the onset of the Scandian collision⁹. The Scandinavian margin of Baltica subducted at rates of 6–10 cm yr⁻¹ beneath Laurentia at ~430 Myr ago^{25,26}, with some estimates considering the entire Scandian event to have been of relatively short duration, ~40 Myr (ref. 27).

Modelling the crustal thermal regime

For the thermal model we used a modified version of the program Pecube²⁸ to estimate the pressure–temperature–time (P – T – t) path of a particle that is subducted and then exhumed during the collision/subduction event. This is done by solving the transient heat transport (by conduction and advection) equation in three dimensions. Figure 7 shows the P – T – t path (lasting 13 Myr) predicted by Pecube for the above tectonic regime of a particle that (1) was initially positioned at mid-crustal levels (20 km) of the subducting plate, (2) was subducted to a depth of 60 km in 1 Myr, (3) resided at that depth for 2 Myr after ‘detaching’ from the subducting plate (the choice of the duration of this interval is discussed below), and (4) was finally exhumed by moving back up the subduction channel to the

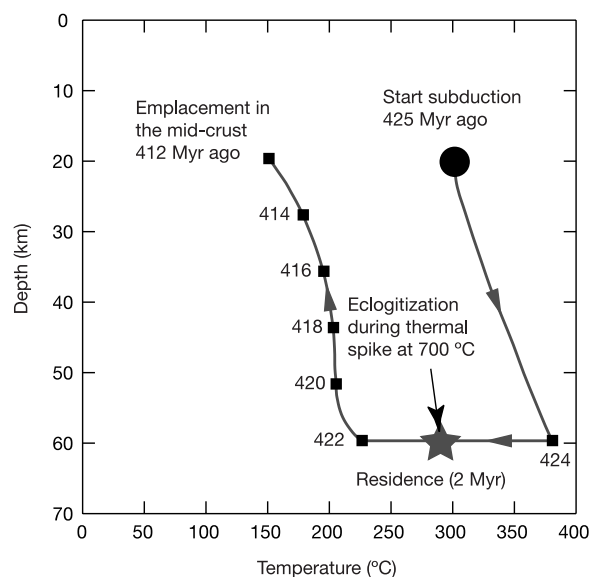


Figure 7 | Modelled pressure–temperature–time path for the Lindås nappe, Bergen arcs. Time markers (square symbols) are at 2 Myr intervals. The following geological constraints were used as input parameters. (1) The granulites of Holsnøy were at depths of 20 km before collision; (2) collision between Baltica and Laurentia commenced 430 Myr ago^{23,44}; (3) subduction 425 Myr ago of the Lindås nappe beneath Laurentia occurred at a rate of 6 cm yr⁻¹ (refs 25, 26); (4) eclogitization during a thermal spike at 700 °C at depths of 60 km (ref. 12) took place 423 Myr ago⁹; (4) exhumation to mid-crustal levels 413 Myr ago is constrained by Rb–Sr mineral isochron ages for amphibolite-facies overprint³⁰. The modelled domain is a 300 km by 100 km cross-section through the upper lithosphere centred on the subduction zone with an angle of 45°. Radiogenic heat production is not considered, and a geothermal gradient of 15 °C km⁻¹ is assumed. The thermal model has three phases: first, subduction for 1 Myr to bury the particle from 20 km to 60 km depth; second, the particle detaches from the subducting plate and resides at 60 km for 2 Myr; third, exhumation to the surface along the subduction zone to the mid-crust in 10 Myr. The residence and exhumation phases of the particle history are related to the formation of a thin channel 5 km wide (the width of Holsnøy island) that forms along the surface of the subducting slab.

mid-crust (20 km) in 10 Myr. Over the entire orogeny, subduction is continuous. The results show that during burial the particle experiences heating to a temperature of 385 °C, followed by substantial cooling after detachment at maximum depth (Fig. 7) because the cold adjacent slab continues to subduct, thus retarding steady-state thermal equilibrium. During exhumation, the particle heats up monotonically, with temperature never exceeding the value reached during burial. Significantly, the thermal modelling shows that with the imposed geological constraints (Fig. 7 legend) the particle never attains temperatures >400 °C.

The integrated thermal spike (total duration of 18 kyr at 526 °C) with concomitant uptake of $^{40}\text{Ar}_\text{E}$ during the period of residence at depth can now be superimposed on the thermal history from Fig. 7, and used in a modified finite-difference numerical model²⁹ of argon mobility in phlogopite. Volume-diffusion modelling using this T - t history yields remarkably good correlations between the calculated diffusion profiles and the age data from the phlogopites (Fig. 5). Although the durations of both subduction (1 Myr) and exhumation (10 Myr) are relatively well-constrained^{4,9,30}, the residence time of the rocks at maximum depth is not well-known. Because of the high sensitivity of the argon systematics to different thermal regimes, however, we can now place narrow limits on the residence time of the subducted crust at maximum depth. Coupling the argon diffusion and thermal models provides an iterative feedback mechanism for the refined fitting of the diffusion profiles (Fig. 5); the results show that residence times longer than 2 Myr result in poor fits.

Spasmodic fluid injection

Finally, by using thermal diffusivity data³¹, we can reconcile the heating of an ultramafic lens with our integrated thermal spike. Assuming an ambient (initial) T of 385 °C (Pecube results) and fluids (passing through the shear zones) acting as a constant heat source (at 700 °C)³², heat-conduction calculations show that the core of a lens 30 m in diameter will attain a temperature of 526 °C in ~10 yr, substantially shorter than our total heating duration estimate of 18 kyr, and at first sight, geologically implausible. However, this apparently contradictory result can be reconciled with our data if the shear zones were repeatedly injected by hot fluid for very short periods (~10 yr), with each injection event only affecting the integrated thermal budget such that the mean temperature of the entire terrane did not exceed 400 °C. This is exactly what would be expected if fluid migration was triggered by multiple, spasmodic deformation events associated with earthquakes. Evidence for earthquakes and brittle deformation abounds in the Bergen arcs, where hydraulic fracturing³³ and the presence of pseudotachylites³⁴ are closely associated with the eclogite-bearing shear zones⁴. Thus, the 18 kyr duration of the thermal spike calculated above represents the integrated time over the 2 Myr residence period of the terrane at 60 km depth in which multiple, short-lived fluid pulses were active.

A new cold-crust model

We therefore propose a new model for the behaviour of the subducting continental lithosphere, in which the rate of subduction and exhumation far outstrips the ability of the temperature of the subducting crust to equilibrate. This alternative interpretation is attractive because it elegantly explains: (1) the widely described 'metastability' of non-eclogitized domains^{2,10,35–37}; (2) steep diffusion profiles in relict minerals in partially transformed eclogite²²; and (3) brittle behaviour of the crust at high pressure³⁴. In this model, the crust remains cool and is repeatedly deformed at high strain rates. Heating is transient, and is caused by either fluid advection along shear zones or frictional heating³⁸. An important consequence of the model is that for transient, episodic fluid flow, the widespread practice of estimating crustal fluid volumes on a time-integrated basis³⁹ does not apply. This study provides a new quantitative approach for constraining the cycle of burial and exhumation of a segment of the crust during orogenesis; given that the bulk of the

crust remained relatively cool (≤ 400 °C), thermal modelling combined with volume-diffusion argon modelling indicate that the timescale for the complete orogenic cycle of the Lindås nappe in the Bergen arcs must have been 13 Myr or less (Fig. 7). This timescale for orogenesis is considerably shorter than previous estimates of ~40 Myr (ref. 27) for the Scandian collision, and may be relevant to continental collisions in general.

Received 15 September 2004; accepted 12 April 2005.

1. Austrheim, H. & Griffin, W. L. Shear deformation and eclogite formation within granulite facies anorthosites of the Bergen Arcs, western Norway. *Chem. Geol.* **50**, 267–281 (1985).
2. Koons, P. O., Rubie, D. C. & Frueh-Green, G. The effects of disequilibrium and deformation on the mineralogical evolution of quartz-diorite during metamorphism in the eclogite facies. *J. Petrol.* **28**, 679–700 (1987).
3. Leech, M. L. Arrested orogenic development: eclogitization, delamination and tectonic collapse. *Earth Planet. Sci. Lett.* **185**, 149–159 (2001).
4. Bjørnerud, M., Austrheim, H. & Lund, M. G. Processes leading to eclogitization (densification) of subducted and tectonically buried crust. *J. Geophys. Res.* **107**(B10), 2252–2269 (2002).
5. Chopin, C. Very-high pressure metamorphism in the western Alps: implications for subduction of continental crust. *Phil. Trans. R. Soc. Lond. A* **321**, 183–197 (1987).
6. Andersen, T. B., Jamtveit, B., Dewey, J. F. & Swensson, E. Subduction and exhumation of continental crust: major mechanisms during continent-continent collision and orogenic extensional collapse, a model based on the south Norwegian Caledonides. *Terra Nova* **3**, 303–310 (1991).
7. Dewey, J. F., Ryan, P. D. & Andersen, T. B. Orogenic uplift and collapse, crustal thickness, fabrics and metamorphic phase changes: the role of eclogites. *Geol. Soc. Lond. Spec. Publ.* **76**, 325–343 (1993).
8. Hynes, A., Arkani-Hamed, J. & Greiling, R. Subduction of continental margins and the uplift of high-pressure metamorphic rocks. *Earth Planet. Sci. Lett.* **140**, 13–25 (1996).
9. Bingen, B., Austrheim, H., Whitehouse, M. J. & Davis, W. J. Trace element signature and U-Pb geochronology of eclogite-facies zircon, Bergen Arcs, Caledonides of W Norway. *Contrib. Mineral. Petrol.* **147**, 671–683 (2004).
10. Jamtveit, B., Bucher-Nurminen, K. & Austrheim, H. Fluid controlled eclogitization of eclogites in deep crustal shear zones, Bergen Arcs, western Norway. *Contrib. Mineral. Petrol.* **104**, 184–193 (1990).
11. Boundy, T. M. & Fountain, D. M. Structural development and petrofabrics of eclogite facies shear zones, Bergen Arcs, western Norway: implications for deep crustal deformation processes. *J. Metamorph. Geol.* **10**, 127–146 (1992).
12. Perchuk, A. L. Eclogites of the Bergen Arcs Complex, Norway: petrology and mineral chronometry. *Petrology* **10**, 99–118 (2002).
13. Kühn, A., Glodny, J., Iden, K. & Austrheim, H. Retention of Precambrian Rb/Sr phlogopite ages through Caledonian eclogite facies metamorphism, Bergen Arc complex, W-Norway. *Lithos* **51**, 305–330 (2000).
14. Bjørnerud, M. & Austrheim, H. Comment on "Evidence for shear-heating, Musgrave Block, central Australia" by A. Camacho, I. McDougall, R. Armstrong and J. Braun. *J. Struct. Geol.* **24**, 1537–1538 (2002).
15. Bingen, B., Davis, W. J. & Austrheim, H. Zircon U-Pb geochronology in the Bergen arc eclogites and their Proterozoic protoliths, and implications for the pre-Scandian evolution of the Caledonides in western Norway. *Geol. Soc. Am. Bull.* **113**, 640–649 (2001).
16. McDougall, I. & Harrison, T. M. *Geochronology and Thermochronology by the $^{40}\text{Ar}/^{39}\text{Ar}$ Method* (Oxford Univ. Press, New York, 1999).
17. Kelley, S. P., Arnaud, N. O. & Turner, S. P. High spatial resolution $^{40}\text{Ar}/^{39}\text{Ar}$ investigations using an ultra-violet laser probe extraction technique. *Geochim. Cosmochim. Acta* **58**, 3519–3525 (1994).
18. Matthey, D., Jackson, D. H., Harris, N. B. W. & Kelley, S. P. Isotopic constraints on fluid infiltration from an eclogite facies shear zone, Holsenøy, Norway. *J. Metamorph. Geol.* **12**, 311–325 (1994).
19. Boundy, T. M., Hall, C. M., Li, G., Essene, E. J. & Halliday, A. N. Fine-scale isotopic heterogeneities and fluids in the deep crust: a $^{40}\text{Ar}/^{39}\text{Ar}$ laser ablation and TEM study of muscovites from a granulite-eclogite transition zone. *Earth Planet. Sci. Lett.* **148**, 223–242 (1997).
20. Foland, K. A. Limited mobility of argon in a metamorphic terrain. *Geochim. Cosmochim. Acta* **43**, 793–801 (1979).
21. Camacho, A. *An Isotopic Study of Deep-crustal Orogenic Processes: Musgrave Block, Central Australia*. Ph.D thesis, Australian National Univ. (1998).
22. Erambert, M. & Austrheim, H. The effect of fluid and deformation on zoning and inclusion patterns in poly-metamorphic garnets. *Contrib. Mineral. Petrol.* **115**, 204–214 (1993).
23. Gee, D. G. A tectonic model for the central part of Scandinavian Caledonides. *Am. J. Sci.* **A 275**, 468–515 (1975).
24. Roberts, D. The Scandinavian Caledonides: event chronology, palaeogeographic settings and likely modern analogues. *Tectonophysics* **365**, 283–299 (2003).
25. Torsvik, T. H. et al. Continental break-up and collision in the Neoproterozoic and Palaeozoic—A tale of Baltica and Laurentia. *Earth Sci. Rev.* **40**, 229–258 (1996).

26. Dewey, J. F. & Strachan, R. A. Changing Silurian–Devonian relative plate motion in the Caledonides: sinistral transpression to sinistral transtension. *J. Geol. Soc. Lond.* **160**, 219–229 (2003).
27. Krabbendam, M. & Dewey, J. F. in *Continental Transpression and Transtensional Tectonics* (eds Holdsworth, R. E., Strachan, R. A. & Dewey, J. F.) 159–181 (Special Publications, Geological Society, London, 1998).
28. Braun, J. Pecube: a new finite-element code to solve the 3D heat transport equation including the effects of a time-varying, finite amplitude surface topography. *Comput. Geosci.* **29**, 787–794 (2003).
29. Lee, J. K. W. Multipath diffusion in geochronology. *Contrib. Mineral. Petrol.* **120**, 60–82 (1995).
30. Glodny, J., Kühn, A. & Austrheim, H. Rb/Sr record of fluid-rock intercation in eclogites, Bergen Arcs, Norway. *Geochim. Cosmochim. Acta* **66** (Suppl. 1), A280 (2002).
31. Seipold, U. Temperature dependence of thermal transport properties of crystalline rocks; a general law. *Tectonophysics* **291**, 161–171 (1998).
32. Seipold, U. & Huenges, E. Thermal properties of gneisses and amphibolites—high pressure and high temperature investigations of KTB-rock samples. *Tectonophysics* **291**, 173–178 (1998).
33. Jamtveit, B., Austrheim, H. & Maltre-Sørensen, A. Accelerated hydration of the Earth's deep crust induced by stress perturbations. *Nature* **408**, 75–78 (2000).
34. Austrheim, H. & Boundy, T. M. Pseudotachylites generated during seismic faulting and eclogitization of the deep crust. *Science* **265**, 82–83 (1994).
35. Wayte, G. J., Worden, R. H., Rubie, D. C. & Droop, G. T. R. A TEM study of disequilibrium plagioclase breakdown at high pressure: the role of infiltrating fluid. *Contrib. Mineral. Petrol.* **101**, 426–437 (1989).
36. Camacho, A. & McDougall, I. Intracratonic, strike-slip partitioned transpression and the formation and exhumation of eclogite facies rocks: An example from the Musgrave Block, central Australia. *Tectonics* **19**, 978–996 (2000).
37. Wain, A. L., Waters, D. J. & Austrheim, H. Metastability of granulites and processes of eclogitisation in the UHP region of western Norway. *J. Metamorph. Geol.* **19**, 609–625 (2001).
38. Camacho, A., McDougall, I., Armstrong, R. & Braun, J. Evidence for shear heating, Musgrave Block, central Australia. *J. Struct. Geol.* **23**, 1007–1013 (2001).
39. Phillipot, P. & Rumble, D. Fluid-rock interactions during high-pressure and ultrahigh-pressure metamorphism. *Int. Geol. Rev.* **42**, 312–327 (2000).
40. Austrheim, H., Erambert, M. & Boundy, T. M. Garnets recording deep crustal earthquakes. *Earth Planet. Sci. Lett.* **139**, 223–238 (1996).
41. Boundy, T. M., Mezger, K. & Essene, E. J. Temporal and tectonic evolution of the granulite-eclogite association from the Bergen Arcs, western Norway. *Lithos* **39**, 159–178 (1997).
42. Giletti, B. J. in *Geochemical Transport and Kinetics* (eds Hofmann, A. W., Giletti, B. J., Yoder, H. S. & Yund, R. A.) 107–115 (Carnegie Institute of Washington, 1974).
43. Harrison, T. M. Diffusion of ^{40}Ar in hornblende. *Contrib. Mineral. Petrol.* **78**, 324–331 (1981).
44. Cocks, L. R. M. & Torsvik, T. H. Earth geography from 500 to 400 million years ago: a faunal and palaeomagnetic review. *J. Geol. Soc. Lond.* **159**, 631–644 (2002).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We especially thank H. Austrheim for all of his help in the field, hospitality, and discussions about the outcrop; M. Villeneuve for use of the ultraviolet-laser argon facility at the Geological Survey of Canada in Ottawa; and in particular S. Smith for technical assistance. In addition, M. Lund helped collect some samples and supplied Fig. 1, and S. Kelley and A. Perchuk provided comments on the manuscript. Comments by H. Austrheim, D. M. Carmichael, A. Clark, L. Godin, I. Parsons, C. Thompson, M. Villeneuve, S. M. Rigden and H. M. Klaschka on earlier versions of this paper are also acknowledged. This research was supported by the Natural Sciences and Engineering Research Council of Canada and the Australian Research Council.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to J.K.W.L. (lee@geol.queensu.ca).

Structure of a Na^+/H^+ antiporter and insights into mechanism of action and regulation by pH

Carola Hunte¹, Emanuela Screpanti¹, Miro Venturi¹, Abraham Rimón², Etana Padan² & Hartmut Michel¹

The control by Na^+/H^+ antiporters of sodium/proton concentration and cell volume is crucial for the viability of all cells. Adaptation to high salinity and/or extreme pH in plants and bacteria or in human heart muscles requires the action of Na^+/H^+ antiporters. Their activity is tightly controlled by pH. Here we present the crystal structure of pH-downregulated NhaA, the main antiporter of *Escherichia coli* and many enterobacteria. A negatively charged ion funnel opens to the cytoplasm and ends in the middle of the membrane at the putative ion-binding site. There, a unique assembly of two pairs of short helices connected by crossed, extended chains creates a balanced electrostatic environment. We propose that the binding of charged substrates causes an electric imbalance, inducing movements, that permit a rapid alternating-access mechanism. This ion-exchange machinery is regulated by a conformational change elicited by a pH signal perceived at the entry to the cytoplasmic funnel.

Regulation of intracellular pH, cellular Na^+ content and cell volume are processes that are essential for all living cells. Na^+/H^+ antiporters¹ have primary functions in these crucial processes^{2–4}. They are integral membrane proteins that are ubiquitous throughout all biological kingdoms; a change in their activity can have an immediate impact on cell metabolism and viability. For example, overactivation of the NHE1 antiporter in heart muscle cells during open-heart surgery in humans has harmful consequences, which are reduced by the use of drugs inhibiting the antiporter^{5,6}. Deletion of plant genes encoding the vacuolar and cytoplasmic membrane Na^+/H^+ antiporters decreases the plant's salt tolerance^{7,8}, whereas overexpressing is exploited to produce salt-resistant plants⁹. Although belonging to different protein families², many prokaryotic Na^+/H^+ antiporters confer on their host cells Na^+ tolerance and/or a capacity to grow at alkaline pH (refs 2, 10). Na^+/H^+ antiporters have also been implicated in the virulence and/or epidemiology of pathogenic enterobacteria¹¹.

NhaA is the main Na^+/H^+ antiporter of *Escherichia coli* and many other enterobacteria¹². Its orthologues are widespread in many other prokaryotes including Archaea. NhaA uses the electrochemical proton gradient maintained across the bacterial membrane and excretes Na^+ in exchange for a 'downhill' flow of protons into the cell¹². NhaA activity is strictly regulated by pH, a property it shares with many other prokaryotic^{12,13} and eukaryotic^{14,15} antiporters and which is essential for cytoplasmic pH regulation. At acidic pH NhaA is downregulated^{12,16}, a feature that permits the formation of crystals ordered well enough to determine its architecture and to reveal the structural basis for a mechanism of Na^+/H^+ exchange and its unique regulation by pH. The structure of NhaA represents a previously undescribed fold that contributes to the understanding of the architecture of membrane proteins in general.

Structure determination

X-ray diffraction data were collected from native and seleno-L-methionine (SeMet)-labelled crystals grown at pH 4.0. Statistics for

data collection and structure determination are summarized in Supplementary Table S1. NhaA crystallized in space group $P2_12_12_1$ with two monomers in the asymmetric unit present in an opposite, non-physiological orientation (Fig. 1a). In the native membrane¹⁷ and in two-dimensional crystals^{18,19}, NhaA is known to form a dimer.

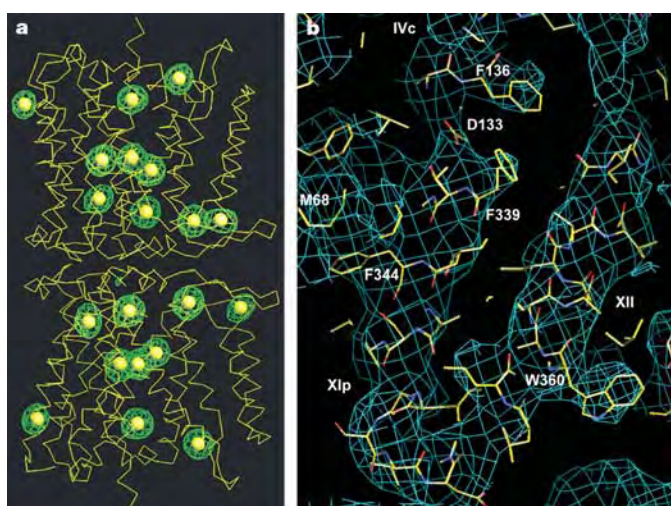


Figure 1 | Experimental electron density. **a**, Side view of the backbone of two NhaA molecules related by two-fold symmetry superimposed with an anomalous Fourier map calculated from the SeMet data set with final phases contoured at 4σ . **b**, Section of the experimental electron-density map (contoured at 1σ) extended to 3.45 Å with the native data and superimposed with the refined NhaA model. Helix XII and the periplasmic helix XIp are connected by a well-defined loop. The short helices XIp (Phe 339) and IVc (Phe 136) are oriented in an anti-parallel manner with residue Asp 133 in an appropriate position for charge compensation.

¹Department of Molecular Membrane Biology, Max Planck Institute of Biophysics, Max-von-Laue-Str. 3, D-60438 Frankfurt, Germany. ²Alexander Silberman Institute of Life Sciences, Hebrew University of Jerusalem, Jerusalem 91904, Israel.

The structure was solved by single-wavelength anomalous dispersion (SAD) taking advantage of 10 SeMet positions per monomer. After solvent flattening and phase extension, the model was built and refined in iterative cycles to a final R_{work} of 29.6% and R_{free} of 31.7% to 3.45 Å resolution. The positions of the SeMet residues were important for model building and sequence assignment. Throughout model building and refinement the presence of a two-fold symmetry was used to constrain the atomic coordinates, and to improve refinement and electron-density map calculations. Experimental electron-density maps are shown in Fig. 1.

Overall architecture

NhaA consists of 388 amino-acid residues with the amino terminus and the carboxy terminus exposed to the cytoplasm²⁰. Our structural model comprises residues 9–384, which are arranged in 12 trans-membrane segments (TMSs; Fig. 2a), in line with previous results^{19,20}. The molecule measures about 40 Å × 45 Å, and its height is about 50 Å. The fold of NhaA is, to the best of our knowledge, previously undescribed; TMSs IV and XI (henceforth designated the TMSs IV/XI assembly) are of opposite orientations in the membrane and each is composed of a short helix, an extended polypeptide chain and a short helix (Fig. 2a, b). The short helices at the periplasmic and cytoplasmic sides are denoted p and c, respectively. In general, it is energetically unfavourable to insert a polar end of a helix in the low-dielectric core of the membrane. Moreover, helices IVc and XIp are oriented in an anti-parallel manner with their partially positive charged N termini facing each other in the middle of the membrane. The position of Asp 133 permits charge compensation and thus stabilization of the arrangement (Figs 1b and 2b). Together these two helices at lower resolution appear misleadingly like a single straight helix, which parallels helix V in close proximity (Fig. 2b). A similar anti-parallel architecture of two short helices with charge compensation exists in the chloride channel²¹. However, whereas in the latter the polypeptide chains turn back, the chains in NhaA cross each other, forming close contacts in the middle of the membrane, and each extends to a second short helix at the respective opposite side of the membrane (helices IVp and XIc, Fig. 2a, b). These latter helices are tilted with respect to helix V and their negative dipole ends are charge compensated by Lys 300 of helix X (Fig. 2b).

Other noteworthy elements are the S-shaped helices III and X, the bent helix IX and the extraordinarily short helices VII and VIII

(Fig. 2a). The membrane boundaries were estimated from the distribution of tryptophan and tyrosine residues²² (Fig. 2a). The periplasmic face of NhaA is flat and formed by structured loops close to the lipid bilayer. The longest loop, loop I–II, consists of a short helix (helix Ia, residues 35–41) followed by an anti-parallel, double-stranded β-sheet (β1, 45–48; β2, 55–58). It is oriented parallel to the membrane, providing a flat and rigid structure at the periplasmic boundary (Fig. 2a). At the cytoplasmic face more flexible loops, and several helices (II, V, IX and XII), protrude into the cytoplasm, creating a rough surface.

The TMSs are organized in two densely packed domains (A and B; Supplementary Fig. S1). Two bundles each containing three TMSs (III, IV and V; X, XI and XII) form domain A. Despite the weak sequence homology the two bundles are structurally related. It is a duplication with opposite orientation with respect to the membrane (Supplementary Fig. S2). Such structural repetitions have been observed in other membrane proteins^{21,23,24}. Domain B is formed by a linear bundle of six helices (II, VII, VIII, IX in the centre, I and VI peripherally attached at opposite sides). The two domains are in full agreement with the organization of helical bundles observed in the three-dimensional map of NhaA obtained by cryo-electron microscopy of two-dimensional crystals, which were also obtained at acidic pH (ref. 19).

At the centre of the domain interface a funnel opens to the cytoplasm (Fig. 3a, c, d, and Supplementary Fig. S1). It is formed by helices II and IX of domain B, helix IVc and its extended chain and helix V of domain A (Fig. 3a). The electrostatic potential surface viewed from the cytoplasm highlights strongly negatively charged patches lining the funnel (Fig. 3c, d). The rest of the cytoplasmic membrane surface is positively charged, whereas the periplasmic surface is negatively charged (Fig. 3e), in line with the positive-inside rule²⁵. The funnel is blocked approximately in the middle of the membrane at the crossing of the extended chains of the TMSs IV/XI assembly (Fig. 3a, d). A shallow negatively charged funnel opens to the periplasm and is formed by helices II, VIII and XIp (Fig. 3a, d, e). The cytoplasmic and periplasmic funnels point towards each other but do not form a continuous pore.

Structural basis for passage and binding site of substrates

This structure is a snapshot of the native, acidic pH-locked conformation of NhaA. Nevertheless, the cytoplasmic cation passage and a

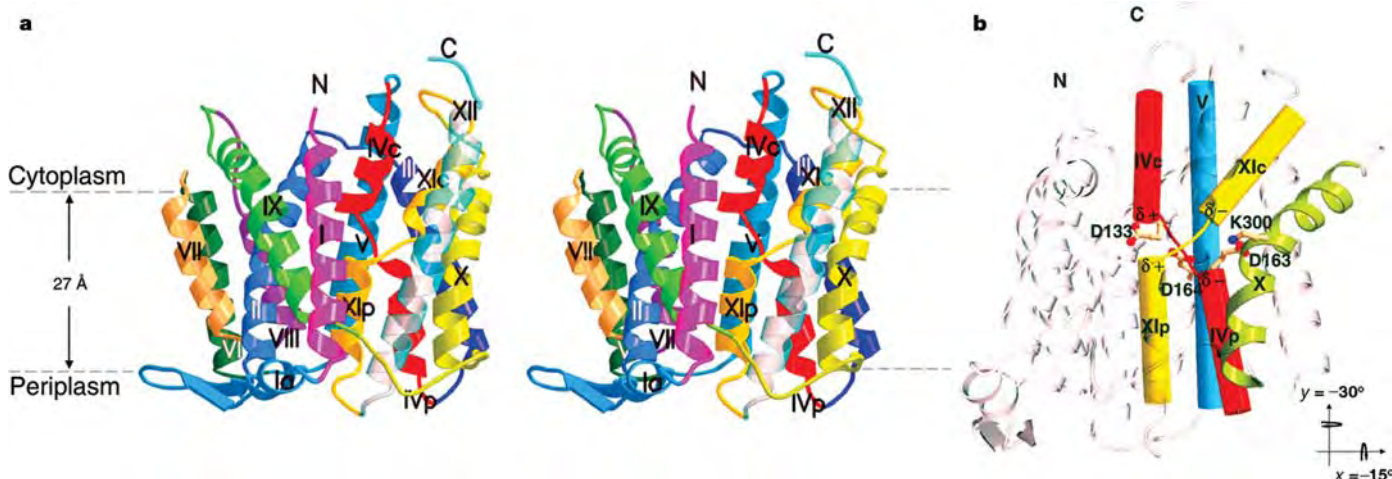


Figure 2 | Overall architecture of NhaA. **a**, Stereo view of a ribbon representation viewed parallel to the membrane (grey broken lines). The 12 TMSs are labelled with roman numerals; they comprise the following residues: 12–30 (I), 59–85 (II), 95–116 (III), 121–131 (IVp), 134–143 (IVc), 150–175 (V), 182–200 (VI), 205–218 (VII), 223–236 (VIII), 247–271 (IX), 290–311 (X), 327–336 (XIc), 340–350 (XIp) and 357–382 (XII). N and C

indicate the N and C termini. **b**, TMSs IV/XI assembly. Helices of the assembly and helix V are shown as cylinders, helix X in ribbon representation. The partial charges of the N and C termini of the short helices are indicated. The orientation of the molecule is indicated with respect to **a**.

most probable binding site for Na^+ and Li^+ , the specific substrates of NhaA, are clearly evident in the structure. The negatively charged funnel is very well suited to attract cations and to increase their local concentrations over that of anions (Fig. 3a). A cluster of three negatively charged residues, Glu 78, Glu 82 and Glu 252, is located on one side of the funnel entrance and residue Asp 11 is on the opposite side (Fig. 3a). An accessibility analysis of the cavity for ions and water with the use of the respective ion radii as the probe size (3.58, 3.82 and 3.31 Å for hydrated Na^+ , Li^+ and K^+ , respectively; 0.95 and 0.6 Å for non-hydrated Na^+ and Li^+ , respectively, and 1.4 Å for water) was performed. The funnel is not selective for cations above the level of the four acidic residues (Glu 78, Glu 82, Glu 252 and Asp 11), and hydrated Na^+ , Li^+ and K^+ as well as water can diffuse into it (Fig. 3a). Subsequently, the passage narrows and is lined by nonpolar residues. Fully hydrated Na^+ or Li^+ ions cannot access the end of the funnel in the middle of the membrane, where Asp 164 of helix V is positioned (Fig. 3a). Although non-hydrated substrates could in principle reach this acidic residue, the structure does not provide a clue as to whether dehydration of the cations occurs in this conformation of NhaA.

The specific location of Asp 164 strongly suggests that it contributes to the Na^+/Li^+ -binding site (Fig. 3a, b, d). Structures of Na^+ -binding proteins demonstrate that the smallest alkali-metal ions are ligated by oxygen atoms provided by water as well as main-chain carbonyl, carboxyl or hydroxyl groups with an average coordination number of five to six (refs 26, 27). Further residues are therefore expected to contribute to the putative Na^+/Li^+ -binding site of NhaA. Indeed, additional polar and/or ionizable amino-acid residues are located near Asp 164 in the centre of the TMSs assembly IV/XI, namely Asp 163 of helix V, and Asp 133 and Thr 132 of TMS IV; they are excluded from the ion passage (Figs 3a, b and 4). Biochemical and

genetic data obtained with NhaA activated at alkaline pH provide a strong clue to the role of these residues in the active conformation. Both Asp 164 and Asp 163 are highly conserved residues and are essential, which is in line with their proposed role in ion binding²⁸. Residues Asp 133 and Thr 132 are also highly conserved but are not essential. However, a mutational change in the two latter residues markedly increased the apparent K_m of the antiporter to both Na^+ and Li^+ , indicating their possible structural role and/or involvement in the translocation mechanism²⁹. It should be noted that Na^+ is not observed in the structure. This might be due to the medium resolution and/or to the partly exposed binding site seen in the acidic-pH-locked conformation of NhaA.

At the periplasmic side, Asp 65 of helix II is located at the tip of the shallow funnel, 16 Å away from Asp 164 (Fig. 3a, d). This funnel is separated from the putative Na^+/Li^+ -binding site by densely packed non-polar residues of helices II, IVp and XIp, which are not accessible to either water or ions, thus forming the periplasmic barrier (Fig. 3a). Hence, the most evident features of the acidic pH-locked structure are the following: first, the cytoplasmic ion passage ending with a narrow cavity that does not allow fully hydrated Na^+ and Li^+ to access the binding site; second, the putative binding site in the centre of the membrane, of which only Asp 164 is exposed to the cytoplasmic ion passage; and last, the barrier at the periplasmic side, to which the TMSs IV/XI assembly contributes.

Mechanism of the pH-regulated Na^+/H^+ exchange

The main physiological role of NhaA is the regulation of cytoplasmic pH and Na^+ (Li^+) content. To maintain a constant intracellular pH of about 7.6 at more alkaline extracellular pH, surplus Na^+ (Li^+), which is toxic to the cell, is excreted in exchange for protons¹². NhaA is tightly regulated by pH. Its activity alters by more than three orders

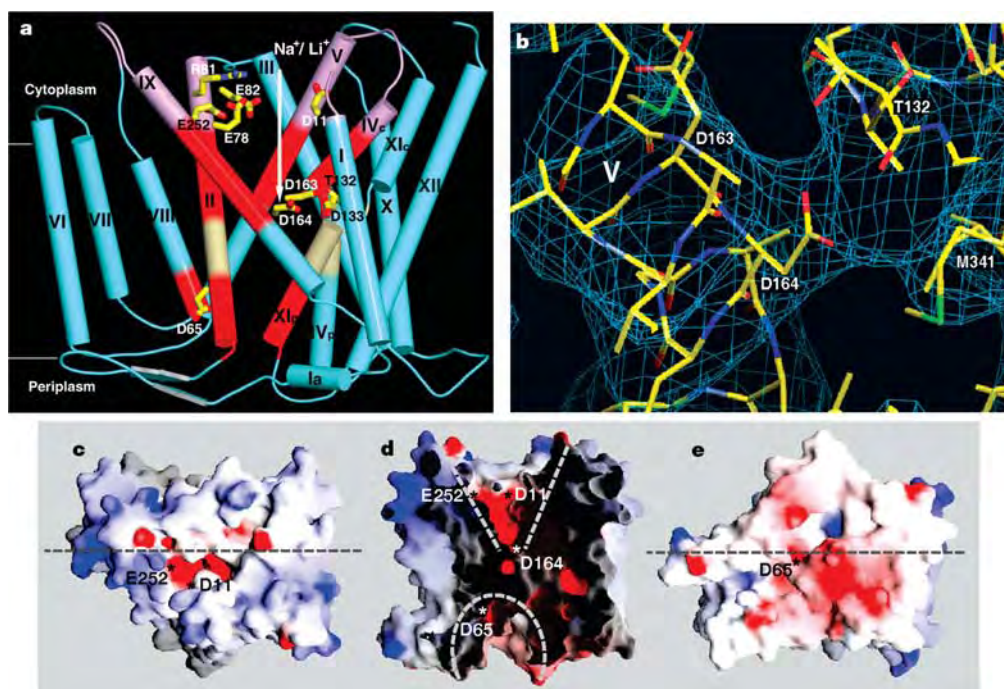


Figure 3 | Substrate passage and periplasmic barrier. **a**, The TMSs lining the cytoplasmic and the periplasmic cation passage are highlighted by colours in the cylinder representation. The cavities were analysed for ion and water accessibility by using the programs CastP and Voidoo^{49,50}. Fully hydrated Na^+/Li^+ can diffuse into the upper zone of the cytoplasmic funnel (light purple) but can enter neither the lower part of it (red) nor the periplasmic funnel (red). The barrier between the passages is coloured in cream. **b**, The putative catalytic site of NhaA viewed from the cytoplasm along the membrane normal. The electron-density map ($2F_o - F_c$)

contoured at 1σ) superimposed with the refined model is shown. **c–e**, The electrostatic potential surface coloured according to its charge (positive in blue, negative in red). Some residue positions (asterisks) are indicated for orientation. **c**, Cytoplasmic view along the membrane normal. **d**, Cross-section through the antiporter with the front part removed along the membrane normal. The position of the section is indicated in **c** and **e** (horizontal line). Cytoplasmic (top) and periplasmic (bottom) funnels are marked by broken lines. **e**, Periplasmic view along the membrane normal.

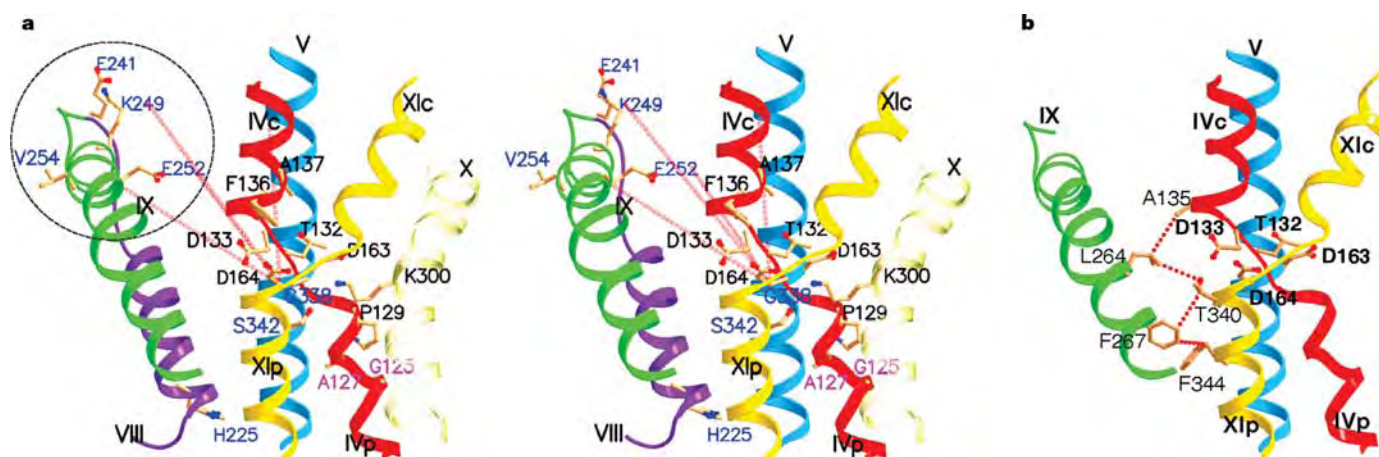


Figure 4 | Structural basis of Na^+/H^+ translocation and pH regulation. **a**, Stereo view of TMSs oriented parallel to the membrane, with the cytoplasmic side at the top. The colour code is as in Fig. 2. Residues whose alterations affect pH regulation^{12,16,31}, apparent K_m (refs 29, 38) or both^{29,38} are shown with side chains and labelled blue, black and purple, respectively.

The putative 'pH sensor' is encircled. The approximate position of the cytoplasmic passage is marked by red dotted lines connecting Asp 164 with residues lining the funnel entry. **b**, Close interactions between helix IX and helices IVc and XIc that are important for pH regulation. Van der Waals contacts between side chains are indicated with dotted red lines.

of magnitude between pH 7 and 8 and it is fully downregulated below pH 6.5 (refs 12, 13). Several questions can be raised about the structure of the acidic pH-locked NhaA. Where is the 'pH sensor'? What is the structural element that transmits and converts the pH signal into a change in activity? How is the change accomplished? The structure reveals that the NhaA N terminus and residue Glu 252 of helix IX at the cytoplasmic funnel entrance are adjacent (Figs 2a and 3a), in line with previous crosslinking data^{30,31}. Mutagenesis of amino-acid residues clustered in these locations and loop VIII–IX drastically affects the pH dependence of NhaA^{16,31,32} (Fig. 4a). Furthermore, conformational changes induced by alkaline pH occur in these locations as probed by accessibility tests: at the N terminus with a monoclonal antibody³², at the N terminus of helix IX with trypsin¹⁶ (Lys 249; Fig. 4a) or with a fluorescent probe³¹ (E252C; Fig. 4a). Most importantly, the pH dependence of these conformational changes at the cytoplasmic side parallels the pH dependence of NhaA activation¹². We therefore suggest that a 'pH sensor', in which a pH signal results in alteration of the protonation state, is located at the entrance of the cytoplasmic passage (Fig. 4a) and that it elicits conformational changes that are transmitted to activate NhaA. In agreement with this, the pH signal is perceived in cells from the cytoplasm³³.

Helix IX is the most likely structural element capable of transmitting the pH signal required to activate NhaA (Fig. 4). It is distorted, a feature that allows flexibility for a long-range conformational change. Its N terminus contributes to the 'pH sensor' and it itself undergoes a pH-induced conformational change^{16,31}. Close to its kink at the centre of the membrane it is in direct contact with the TMSs IV/XI assembly (Fig. 4b), an essential part of the Na^+/H^+ exchange machinery (see below).

The generally accepted model explaining Na^+/H^+ exchange³⁴ and many other secondary transport processes^{35,36} is the alternating-access mechanism. In this model, the transporter has two major alternating conformations with the substrate-binding site facing either inwards or outwards. Interconversion between the two conformations in an antiporter is only possible through a substrate-bound form of the protein. For NhaA that means that either one Na^+/Li^+ or two H^+ are bound³⁷. The two aspartates of the binding site are located on the straight helix V, which does not have conspicuous structural features for conformational changes. In contrast, the TMSs IV/XI assembly, with its short helices and their dipoles compensated for in the middle of the membrane by Asp 133 and Lys 300, is most suitable for charge-induced subtle and fast

conformational changes. On the basis of these unique structural features we suggest that a pH-regulated movement of the extended chains of the TMSs IV/XI assembly allows alternating access of the substrate-binding site to either the intracellular or extracellular space (Fig. 5).

At acidic pH, when it is downregulated, only part of the cation-binding site (Asp 164) is exposed to the cation passage while the periplasmic passage is blocked by the ion-barrier formed partly by helix XIc (Figs 3a and 5a). We suggest that at alkaline pH, in response to a signal from the 'pH sensor', a conformational change of helix IX results in the reorientation of helices XIc and IVc. This would expose the full Na^+/Li^+ -binding site (Asp 164, Asp 163 and most probably Thr 132) to the cytoplasmic passage and remove the periplasmic ion barrier, leaving the extended chain to seal the binding site from the periplasm (Fig. 5b). This conformation would now be ready for Na^+/H^+ exchange. Binding of Na^+/Li^+ to the active site from the cytoplasm would result in a charge imbalance that then would trigger a small movement of XIc and IVc and their extended chains. The cation-loaded binding site would thus then be exposed to the periplasm and sealed from the cytoplasm (Fig. 5c). On release of Na^+/Li^+ , both aspartates would become protonated, thereby inducing a conformational change that would expose them back to the cytoplasm; deprotonation would complete the cycle (Fig. 5b). This model permits an overall stoichiometry of one Na^+/Li^+ exported in exchange for two protons taken up. Because NhaA is reversible, the magnitude of the electrochemical potential difference of Na^+/Li^+ compared with that of H^+ determines the direction of the cation exchange across the membrane. The rate of exchange would be determined by the rate of movement of the extended chains and helices IVc and XIc in response to pH. Acidic pH locks NhaA, whereas alkaline pH activates it (Fig. 5a).

This working model is strongly supported by genetic and biochemical data obtained at alkaline pH, at which NhaA is active. The measured $\text{Na}^+:\text{H}^+$ stoichiometry³⁷ of the exchange is 1:2. Many residues in TMS IV are conserved in the NhaA protein family, and their mutations markedly affect³⁸ the apparent K_m for Na^+ and Li^+ (Fig. 4a). Furthermore, the need for the reorientation of helix IVc with respect to XIc is strongly supported by the double mutant F136C/S342C (ref. 38). This mutant forms a disulphide bond that inhibits transport, and reducing the bond restores activity in a reversible manner³⁸. Extensive crosslinking data are in accordance with the close contacts between the TMSs of the assembly³⁸ and their crucial role for activity and pH regulation^{29,38}. The most informative

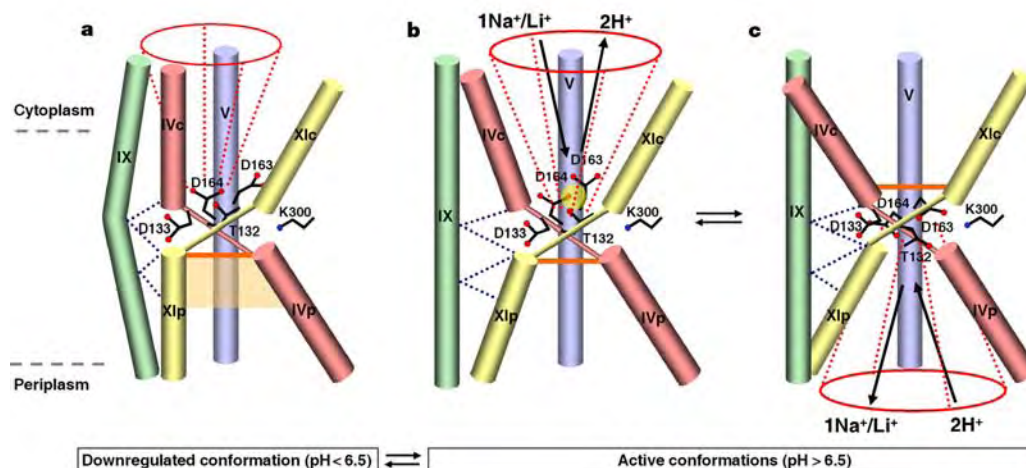


Figure 5 | Proposed mechanism of pH regulation and translocation of NhaA. The TMSs IV/XI assembly, the charge-compensating residues Asp 133 and Lys 300, and further structural elements involved are shown schematically. **a**, Acidic pH-locked conformation. Ion transport is prevented by the periplasmic ion barrier (transparent, cream-coloured area) and by the only partly exposed $\text{Na}^+(\text{Li}^+)$ -binding site (residues Asp 164, Asp 163 and Thr 132). **b**, Activation by alkaline pH induces conformational changes in helix IX resulting in the reorientation of helices IVc and XIp (interactions

indicated by blue dotted lines). The putative $\text{Na}^+(\text{Li}^+)$ -binding site (yellow transparent circle) is now exposed to the cytoplasmic funnel (red dotted lines and red circle) and sealed towards the periplasm (orange bar). **c**, $\text{Na}^+(\text{Li}^+)$ binding results in the opening of the periplasmic funnel and the exposure of the active site to the periplasm. The cation is released. Protonation of the aspartates (Asp 164 and Asp 163) brings the antiporter back to the active conformation open to the cytoplasm.

examples are the mutants G338S (ref. 39) and G338C (ref. 38) of the extended chain XI (Fig. 4a) that completely remove the pH control and produce NhaA variants that are fully active independently of pH. In the inactive double variant T132C/G338C both activity and pH control can be partly restored by re-establishing the physical interaction between the extended chains by chemical crosslinking³⁸.

Discussion

The physical separation between the 'pH sensor' and the exchange machinery revealed by the structure entails long-range, pH-induced conformational changes for pH activation. This observation is in agreement with biochemical data obtained from both prokaryotic¹² and eukaryotic Na^+/H^+ antiporters^{40,41}. The suggested mechanism of Na^+/H^+ exchange of the alkaline pH-activated NhaA is based on small conformational changes limited to a small part of the protein. In contrast, the structures of four secondary transporters^{42–45} have recently indicated a different structural basis for the transport mechanism. Heavily distorted helices with kinks and bends lining a wide substrate passage provide structural flexibility, thereby allowing the molecule to assume different conformations. This type of mechanism is suitable for these slow transporters of large organic molecules with catalytic-centre activities (turnover numbers) of 600, 800 and $1,440 \text{ min}^{-1}$ (LacY (ref. 46), ADP/ATP antiporter⁴⁷ and GlpT antiporter⁴⁸, respectively). In contrast, NhaA is one of the fastest transporters known¹³, with a catalytic-centre activity of $89,000 \text{ min}^{-1}$. The suggested mechanism of Na^+/H^+ exchange, based on subtle conformational changes in the centre of the molecule, is therefore most suitable for the very high activity of NhaA.

Although many aspects of the ion-translocation mechanism and pH regulation are illuminated by this NhaA structure, higher-resolution structural determinations with Li^+ or Na^+ bound are required to understand the ligand binding and translocation mechanism at the atomic level. The alkaline pH-induced conformation is essential for a further understanding of the pH control and proton access to the binding site.

METHODS

Overexpression and purification. Cells of *E. coli* RK20 were co-transformed with the plasmids pAXH and pI^Q encoding His-tagged NhaA and LaCl_3 ,

respectively. The gene *nhaA*, which was deleted from strain RK20, was expressed from the plasmid after induction with isopropyl- β -D-thiogalactoside. Membranes were prepared from broken cells by differential centrifugation. The protein was solubilized in 1% n-dodecyl- β -D-maltopyranoside solution and purified on a nickel-affinity column. The detergent was exchanged on the column with 0.03% n-dodecyl- α -D-maltopyranoside. The eluted protein was concentrated and supplemented with 20% w/v sucrose. SeMet-labelled protein was produced and purified accordingly, but with *E. coli* B834DE3 (Novagen) cells were grown in minimal medium containing seleno-L-methionine.

Crystallization. Crystals were grown at 6°C by hanging-drop vapour diffusion by mixing equal volumes of protein ($4\text{--}6 \text{ mg ml}^{-1}$) and reservoir solution (28–34% v/v poly(ethylene glycol) 400, 200–450 mM MgCl_2 , 100 mM KCl, 25 mM sodium citrate pH 4, 1% n-octyl- β -D-glucopyranoside and 0.5% ethanol). Crystals were flash-cooled to 100 K with nitrogen.

Structure determination and refinement. Diffraction data were collected at the ESRF beam lines ID29 (data set SeMet SAD, $4.3 \text{ \AA}$ resolution) and ID23 (data set Native 1, $3.8 \text{ \AA}$ resolution) and at the SLS beam line PX06 (data set Native 2, $3.45 \text{ \AA}$ resolution) with charge-coupled device detectors and cryo-cooling at 100 K (Supplementary Table S1). The structure was solved by SAD, locating 20 selenium atoms. Phases were calculated to $4.3 \text{ \AA}$ ($17 \text{ \AA}$ – $4.3 \text{ \AA}$) with a figure of merit of 0.42 for centric reflections and 0.13 for centric reflections. The SAD phases were applied to the native data set (Native 1) and gradually extended to $3.8 \text{ \AA}$ with solvent flattening, histogram matching and averaging options. Model building and refinement were facilitated by locating the selenium atom positions in anomalous difference maps from the SeMet data (Fig. 1a) and by keeping a tight two-fold non-crystallographic symmetry restraint during refinement. For final refinement of the model the second native data set at $3.45 \text{ \AA}$ was used.

Further information. Detailed information on methodological aspects is provided as Supplementary Information.

Received 20 October 2004; accepted 22 April 2005.

- West, I. & Mitchell, P. Proton/sodium ion antiport in *Escherichia coli*. *Biochem. J.* **144**, 87–90 (1974).
- Padan, E., Venturi, M., Gerchman, Y. & Dover, N. Na^+/H^+ antiporters. *Biochim. Biophys. Acta* **1505**, 144–157 (2001).
- Orlowski, J. & Grinstein, S. Diversity of the mammalian sodium/proton exchanger SLC9 gene family. *Pflugers Arch.* **447**, 549–565 (2004).
- Counillon, L. & Pouyssegur, J. The expanding family of eucaryotic Na^+/H^+ exchangers. *J. Biol. Chem.* **275**, 1–4 (2000).
- Engelhardt, S., Hein, L., Keller, U., Klambt, K. & Lohse, M. J. Inhibition of Na^+/H^+ exchange prevents hypertrophy, fibrosis, and heart failure in β 1-adrenergic receptor transgenic mice. *Circ. Res.* **90**, 814–819 (2002).
- Myers, M. L. in *The Sodium-Hydrogen Exchange, from Molecule to its Role in Disease* (eds Karmazyn, N., Avkiran, M. & Fliegel, L.) 279–290 (Kluwer Academic, Boston, Massachusetts, 2003).

7. Apse, M. P., Sottosanto, J. B. & Blumwald, E. Vacuolar cation/H⁺ exchange, ion homeostasis, and leaf development are altered in a T-DNA insertional mutant of AtNHX1, the *Arabidopsis* vacuolar Na⁺/H⁺ antiporter. *Plant J.* **36**, 229–239 (2003).
8. Qiu, Q. S., Barkla, B. J., Vera-Estrella, R., Zhu, J. K. & Schumaker, K. S. Na⁺/H⁺ exchange activity in the plasma membrane of *Arabidopsis*. *Plant Physiol.* **132**, 1041–1052 (2003).
9. Apse, M. P., Aharon, G. S., Snedden, W. A. & Blumwald, E. Salt tolerance conferred by overexpression of a vacuolar Na⁺/H⁺ antiport in *Arabidopsis*. *Science* **285**, 1256–1258 (1999).
10. Krulwich, T. A., Ito, M. & Guffanti, A. A. The Na⁺-dependence of alkaliphily in *Bacillus*. *Biochim. Biophys. Acta* **1505**, 158–168 (2001).
11. Hase, C. C. & Mekalanos, J. J. Effects of changes in membrane sodium flux on virulence gene expression in *Vibrio cholerae*. *Proc. Natl Acad. Sci. USA* **96**, 3183–3187 (1999).
12. Padan, E., Tzuber, T., Herz, K., Kozachkov, L. & Galili, L. NhaA of *Escherichia coli*, as a model of a pH regulated Na⁺/H⁺ antiporter. *Biochim. Biophys. Acta* **1658**, 2–13 (2004).
13. Taglicht, D., Padan, E. & Schuldiner, S. Overproduction and purification of a functional Na⁺/H⁺ antiporter coded by *nhaA* (ant) from *Escherichia coli*. *J. Biol. Chem.* **266**, 11289–11294 (1991).
14. Putney, L. K., Denker, S. P. & Barber, D. L. The changing face of the Na⁺/H⁺ exchanger, NHE1: structure, regulation and cellular actions. *Annu. Rev. Pharmacol. Toxicol.* **42**, 527–552 (2002).
15. Wakabayashi, S., Pang, T., Hisamitsu, T. & Shigekawa, M. in *The Sodium–Hydrogen Exchange, from Molecule to its Role in Disease* (eds Karmazyn, N., Avkiran, M. & Fliegel, L.) 35–49 (Kluwer Academic, Boston, Massachusetts, 2003).
16. Gerchman, Y., Rimon, A. & Padan, E. A pH-dependent conformational change of NhaA Na⁺/H⁺ antiporter of *Escherichia coli* involves loop VIII–IX, plays a role in the pH response of the protein, and is maintained by the pure protein in dodecyl maltoside. *J. Biol. Chem.* **274**, 24617–24624 (1999).
17. Gerchman, Y., Rimon, A., Venturi, M. & Padan, E. Oligomerization of NhaA, the Na⁺/H⁺ antiporter of *Escherichia coli* in the membrane and its functional and structural consequences. *Biochemistry* **40**, 3403–3412 (2001).
18. Williams, K. A., Geldmacher-Kaufner, U., Padan, E., Schuldiner, S. & Kuhlbrandt, W. Projection structure of NhaA, a secondary transporter from *Escherichia coli*, at 4.0 Å resolution. *EMBO J.* **18**, 3558–3563 (1999).
19. Williams, K. A. Three-dimensional structure of the ion-coupled transport protein NhaA. *Nature* **403**, 112–115 (2000).
20. Rothman, A., Padan, E. & Schuldiner, S. Topological analysis of NhaA, a Na⁺/H⁺ antiporter from *Escherichia coli*. *J. Biol. Chem.* **271**, 32288–32292 (1996).
21. Dutzler, R., Campbell, E. B., Cadene, M., Chait, B. T. & MacKinnon, R. X-ray structure of a ClC chloride channel at 3.0 Å reveals the molecular basis of anion selectivity. *Nature* **415**, 287–294 (2002).
22. Palsdottir, H. & Hunte, C. Lipids in membrane protein structures. *Biochim. Biophys. Acta* **1666**, 2–18 (2004).
23. Fu, D. *et al.* Structure of a glycerol-conducting channel and the basis for its selectivity. *Science* **290**, 481–486 (2000).
24. Khademi, S. *et al.* Mechanism of ammonia transport by Amt/MEP/Rh: structure of AmtB at 1.35 Å. *Science* **305**, 1587–1594 (2004).
25. von Heijne, G. Membrane protein structure prediction: hydrophobicity analysis and the 'positive inside' rule. *J. Mol. Biol.* **225**, 487–494 (1992).
26. Glusker, J. P. Structural aspects of metal liganding to functional groups in proteins. *Adv. Protein Chem.* **42**, 1–76 (1991).
27. Harding, M. M. The architecture of metal coordination groups in proteins. *Acta Crystallogr. D Biol. Crystallogr.* **60**, 849–859 (2004).
28. Inoue, H., Noumi, T., Tsuchiya, T. & Kanazawa, H. Essential aspartic acid residues, Asp-133, Asp-163 and Asp-164, in the transmembrane helices of a Na⁺/H⁺ antiporter (NhaA) from *Escherichia coli*. *FEBS Lett.* **363**, 264–268 (1995).
29. Galili, L., Rothman, A., Kozachkov, L., Rimon, A. & Padan, E. Trans membrane domain IV is involved in ion transport activity and pH regulation of the NhaA-Na⁺/H⁺ antiporter of *Escherichia coli*. *Biochemistry* **41**, 609–617 (2002).
30. Rimon, A., Tzuber, T., Galili, L. & Padan, E. Proximity of cytoplasmic and periplasmic loops in NhaA Na⁺/H⁺ antiporter of *Escherichia coli* as determined by site-directed thiol cross-linking. *Biochemistry* **41**, 14897–14905 (2002).
31. Tzuber, T., Rimon, A. & Padan, E. Mutation E252C increases drastically the *K_m* for Na and causes an alkaline shift of the pH dependence of NhaA Na⁺/H⁺ antiporter of *Escherichia coli*. *J. Biol. Chem.* **279**, 3265–3272 (2004).
32. Venturi, M. *et al.* The monoclonal antibody 1F6 identifies a pH-dependent conformational change in the hydrophilic NH₂ terminus of NhaA Na⁺/H⁺ antiporter of *Escherichia coli*. *J. Biol. Chem.* **275**, 4734–4742 (2000).
33. Zilberstein, D., Agmon, V., Schuldiner, S. & Padan, E. The sodium/proton antiporter is part of the pH homeostasis mechanism in *Escherichia coli*. *J. Biol. Chem.* **257**, 3687–3691 (1982).
34. Jin, J., Guffanti, A. A., Bechhofer, D. H. & Krulwich, T. A. Tet(L) and tet(K) tetracycline-divalent metal/H⁺ antiporters: characterization of multiple catalytic modes and a mutagenesis approach to differences in their efflux substrate and coupling ion preferences. *J. Bacteriol.* **184**, 4722–4732 (2002).
35. Jencks, W. P. The utilization of binding energy in coupled vectorial processes. *Adv. Enzymol.* **51**, 75–106 (1980).
36. West, I. C. Ligand conduction and the gated-pore mechanism of transmembrane transport. *Biochim. Biophys. Acta* **1331**, 213–234 (1997).
37. Taglicht, D., Padan, E. & Schuldiner, S. Proton–sodium stoichiometry of NhaA, an electrogenic antiporter from *Escherichia coli*. *J. Biol. Chem.* **268**, 5382–5387 (1993).
38. Galili, L., Hertz, K., Dym, O. & Padan, E. Unraveling functional and structural interactions between trans membrane domains IV and XI of NhaA Na⁺/H⁺ antiporter of *Escherichia coli*. *J. Biol. Chem.* **279**, 23104–23113 (2004).
39. Rimon, A., Gerchman, Y., Kariv, Z. & Padan, E. A point mutation (G338S) and its suppressor mutations affect both the pH response of the NhaA-Na⁺/H⁺ antiporter as well as the growth phenotype of *Escherichia coli*. *J. Biol. Chem.* **273**, 26470–26476 (1998).
40. Aronson, P. Kinetic properties of the plasma membrane Na⁺/H⁺ exchanger. *Annu. Rev. Physiol.* **47**, 545–560 (1985).
41. Wakabayashi, S., Hisamitsu, T., Pang, T. & Shigekawa, M. Kinetic dissection of two distinct proton binding sites in Na⁺/H⁺ exchangers by measurement of reverse mode reaction. *J. Biol. Chem.* **278**, 43580–43585 (2003).
42. Murakami, S., Nakashima, R., Yamashita, E. & Yamaguchi, A. Crystal structure of bacterial multidrug efflux transporter AcrB. *Nature* **419**, 587–593 (2002).
43. Abramson, J. *et al.* Structure and mechanism of the lactose permease of *Escherichia coli*. *Science* **301**, 610–615 (2003).
44. Pebay-Peyroula, E. *et al.* Structure of mitochondrial ADP/ATP carrier in complex with carboxyatractylolide. *Nature* **426**, 39–44 (2003).
45. Huang, Y., Lemieux, M. J., Song, J., Auer, M. & Wang, D. N. Structure and mechanism of the glycerol-3-phosphate transporter from *Escherichia coli*. *Science* **301**, 616–620 (2003).
46. Viitanen, P., Garcia, M. L. & Kaback, H. R. Purified reconstituted lac carrier protein from *Escherichia coli* is fully functional. *Proc. Natl Acad. Sci. USA* **81**, 1629–1633 (1984).
47. Klingenberg, M. in *The ADP/ATP Carrier in Mitochondrial Membranes* (ed. Martonosi, A. N.) (Plenum, New York, 1985).
48. Auer, M. *et al.* High-yield expression and functional analysis of *Escherichia coli* glycerol-3-phosphate transporter. *Biochemistry* **40**, 6628–6635 (2001).
49. Liang, J., Edelsbrunner, H. & Woodward, C. Anatomy of protein pockets and cavities: measurement of binding site geometry and implications for ligand design. *Protein Sci.* **7**, 1884–1897 (1998).
50. Kleywegt, G. J. & Jones, T. A. Detection, delineation, measurement and display of cavities in macromolecular structures. *Acta Crystallogr. D Biol. Crystallogr.* **50**, 178–185 (1994).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We acknowledge the beamtime and the assistance of the personnel at beamlines ID29 and ID23 at the European Synchrotron Facility and at PX06 at the Swiss Light Source. This study was supported by grants from the German Israeli Foundation for Scientific Research and Development (to H.M. and E.P.), the Max Planck Society, the Fonds der Chemischen Industrie and the Israel Science Foundation (to E.P.). E.S. was supported by the International Max Planck Research School (IMPRS).

Author Information Coordinates have been deposited in the Protein Data Bank under accession code 1ZCD. Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to C.H. (carola.hunte@mpibp-frankfurt.mpg.de), E.P. (etana@vms.huji.ac.il) or H.M. (hartmut.michel@mpibp-frankfurt.mpg.de).

The U/Th production ratio and the age of the Milky Way from meteorites and Galactic halo stars

Nicolas Dauphas¹

Some heavy elements (with atomic number $A > 69$) are produced by the 'rapid' (r)-process of nucleosynthesis, where lighter elements are bombarded with a massive flux of neutrons^{1–8}. Although this is characteristic of supernovae and neutron star mergers, uncertainties in where the r-process occurs persist because stellar models are too crude to allow precise quantification of this phenomenon. As a result, there are many uncertainties and assumptions in the models used to calculate the production ratios of actinides (like uranium-238 and thorium-232). Current estimates of the U/Th production ratio range from ~ 0.4 to 0.7 . Here I show that the U/Th abundance ratio in meteorites⁹ can be used, in conjunction with observations of low-metallicity stars in the halo of the Milky Way^{10–12}, to determine the U/Th production ratio very precisely ($0.571^{+0.037}_{-0.031}$). This value can be used in future studies to constrain the possible nuclear mass formulae used in r-process calculations^{5,6}, to help determine the source of Galactic cosmic rays, and to date circumstellar grains⁵. I also estimate the age of the Milky Way ($14.5^{+2.8}_{-2.2}$ Gyr) in a way that is independent of the uncertainties associated with fluctuations in the microwave background¹³ or models of stellar evolution^{14,15}.

Cosmochemistry is the art of dating the duration of nucleosynthesis and the age of the Galaxy using the abundances of radioactive elements in the Solar System^{1–8,16–20}. For all but the most volatile elements, the composition of primitive carbonaceous chondrites is identical to that measured in the solar photosphere²¹. The U/Th ratio measured in these meteorites therefore represents a reliable estimate of the bulk Solar System ratio, which is 0.270 ± 0.004 (ref. 9; note that hereafter all errors given are 1σ or 68% confidence intervals). After correction for decay during the 4.567 Gyr since the condensation of the first solids in the Solar System, this corresponds to an initial ratio of 0.438 ± 0.006 in the interstellar medium at Solar System birth ($R_{\odot}^{U/Th}$). This value is the result of a competition between decay and nucleosynthesis integrated over the presolar history of the Galaxy^{2,4,17–20}, a process that is also known as Galactic chemical evolution (GCE). If the history of nucleosynthesis is specified, and if the age of the Galaxy (T_G) and the U/Th production ratio ($P^{U/Th}$) are known, then it is possible to predict what $R_{\odot}^{U/Th}$ should be. Conversely, if the history of nucleosynthesis is specified and $R_{\odot}^{U/Th}$ is known, then it is possible to derive a relationship between T_G and $P^{U/Th}$.

The simplest GCE is the so-called closed-box model²⁰ (Supplementary Information). It however fails to reproduce important astronomical observations. Notably, it predicts far more low metallicity G-dwarfs than are actually observed^{20,22} (Fig. 1). Various explanations have been advanced to remedy this problem²⁰. The most likely is that the Milky Way did not behave as a closed box, but instead grew from the accretion of matter throughout its history. High velocity clouds containing low metallicity gas ($\sim (0.1–0.2)Z_{\odot}$,

where $Z_{\odot}$ is the solar metallicity) are indeed seen being accreted by the Galaxy²³. If infall of such low metallicity material is taken into account, then it is possible to devise GCE models that reproduce the observed G-dwarf metallicity distribution²⁰. Dauphas *et al.*²⁴ developed an algorithm that allows the calculation, for a given T_G , of the parameters of a nonlinear²⁵ open GCE model that is constrained by, and therefore reproduces, the present gas surface density ($13 \pm 3 M_{\odot} \text{pc}^{-2}$; ref. 26), the present total surface density of the disk ($56 \pm 6 M_{\odot} \text{pc}^{-2}$; ref. 27), the solar metallicity at Solar System birth²¹, and the G-dwarf metallicity distribution²² (see Fig. 1, and Supplementary Information). When these parameters are known, it is straightforward to calculate $P^{U/Th}$ and its uncertainty. This

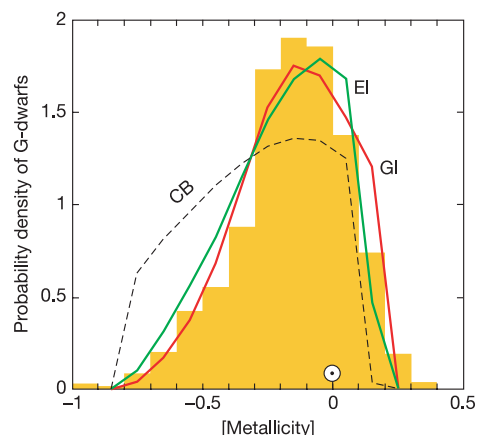


Figure 1 | Distribution of G-dwarf metallicity normalized to solar composition.

G-dwarfs are low mass stars that are not active sites of nucleosynthesis and have lifetimes that exceed the age of the Galaxy. Their compositions can therefore be used to probe the nucleosynthesis of metals and the enrichment of the interstellar medium through time. The yellow histogram is the observed distribution²². The closed-box model of GCE with prompt initial enrichment (black dashed curve, labelled CB) overproduces the number of low metallicity G-dwarfs compared to what is actually observed^{20,22}. GCE models incorporating infall of gas on the disk give better matches to the observed distribution. The infall rate can be parameterized as a gaussian (red curve labelled GI, here for $T_G = 14.5$ Gyr) or an exponential (green curve labelled EI, here for $T_G = 15.1$ Gyr). The rate of infall is parameterized as $\xi f(\tau)$, where ξ is a scaling constant and τ is a time constant. For a gaussian infall, f is a normal distribution with standard deviation equal to the mean, $N(\tau, \tau)$. For an exponential infall, f takes the form $\exp(-t/\tau)$. The rate of star formation is a power function of the gas surface density, $\omega \Sigma_g^n$, where ω is the star formation constant and $n = 1.4$ (ref. 25). For a gaussian infall with $T_G = 14.5$ Gyr, the optimum parameters are $\xi = 66.6 M_{\odot} \text{pc}^{-2} \text{Gyr}^{-1}$, $\tau = 3.46$ Gyr and $\omega = 0.0462 M_{\odot}^{-0.4} \text{pc}^{0.8} \text{Gyr}^{-1}$. For an exponential infall with $T_G = 15.1$ Gyr, the optimum parameters are $\xi = 14.2 M_{\odot} \text{pc}^{-2} \text{Gyr}^{-1}$, $\tau = 4.03$ Gyr and $\omega = 0.0419 M_{\odot}^{-0.4} \text{pc}^{0.8} \text{Gyr}^{-1}$. See Supplementary Information for details.

¹Origins Laboratory, Department of the Geophysical Sciences, Enrico Fermi Institute, and Chicago Center for Cosmochemistry, The University of Chicago, 5734 South Ellis Avenue, Chicago Illinois 60637, USA.

procedure can be repeated for any given age between 10 and 20 Gyr, and the relationship between T_G and $P^{U/Th}$ can be derived. Note that the rate of infall can be parameterized as an exponential²⁰ or as a gaussian²⁸. The differential equations are nonlinear and only a numerical treatment is possible. The result can be approximated by a simple formula, valid between 10 and 20 Gyr ago (details of the mathematical treatment of GCE can be found in Supplementary Information):

$$P^{U/Th} = R_{\odot}^{U/Th} / (aT_G + b) \quad (1)$$

where $a = -1.576 \times 10^{-2}$ and $b = 0.9946$ for a gaussian rate of infall, and $a = -1.968 \times 10^{-2}$ and $b = 1.0114$ for an exponential rate of infall; time is in Gyr.

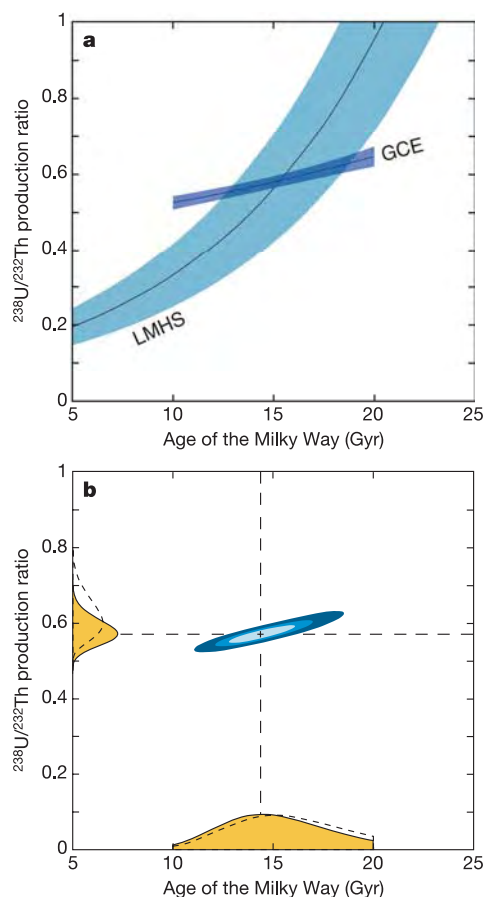


Figure 2 | Determinations of the U/Th production ratio and the age of the Milky Way. **a**, The light blue curve labelled LMHS is derived from the determination of the U/Th abundance ratio in a low metallicity halo star, CS 31082–001 (refs 10,11, equation (2)). The dark blue curve labelled GCE is derived from the solar U/Th ratio⁹ and a GCE model²⁴ incorporating infall of low metallicity gas with a rate parameterized as a gaussian (equation (1)). For a given age of the Milky Way, a GCE model can be built on the basis of observations of the gas and the total surface densities^{26,27} of the Galactic disk, the metallicity of the Sun²¹, and the G-dwarf metallicity distribution²². This GCE model can then be used to calculate the U/Th production ratio required to explain the U/Th ratio measured in meteorites⁹. Repeating this procedure for a range of ages allows construction of the GCE curve. Because the U/Th ratio at Solar System formation integrates nucleosynthesis and decay over Galactic history, the position of the GCE curve is not very sensitive to the details of the GCE model. See Supplementary Information for details. **b**, The contours of the uncertainty ellipsoid of the intersection correspond to 20%, 38% and 68% confidence intervals. The curves filled in yellow are the marginal probability distributions of the U/Th production ratio ($P^{U/Th} = 0.571^{+0.037}_{-0.031}$) and the age of the Milky Way ($T_G = 14.5^{+2.8}_{-2.2}$ Gyr). The dashed curves are the marginal probability distributions if an exponential rate of infall is adopted ($P^{U/Th} = 0.602^{+0.058}_{-0.043}$ and $T_G = 15.1^{+2.8}_{-2.3}$ Gyr). Error bars are 68% confidence intervals.

An additional constraint on the relationship between T_G and $P^{U/Th}$ is given by the recent determination of U and Th abundances in the low metallicity halo stars (LMHS) CS 31082–001 and BD +17°3248, that show pronounced enrichments in the products of r-process nucleosynthesis^{10–12}. The most precise U/Th ratio measured in CS 31082–001 ($R_{LMHS}^{U/Th}$) is 0.115 ± 0.029 (ref. 11). Low metallicity stars formed very early in Galactic history, presumably within 0.1–0.3 Gyr of the formation of the halo^{29,30}. The initial U/Th ratio trapped in CS 31082–001 was produced by an earlier generation of supernovae or neutron star mergers and was only modified by free decay during the life of the star. A second relationship between T_G and $P^{U/Th}$ can thus be derived:

$$P^{U/Th} = R_{LMHS}^{U/Th} e^{(\lambda_U - \lambda_{Th})T_G} \quad (2)$$

where λ_U and λ_{Th} are the decay constants of ^{238}U and ^{232}Th (half-lives of 4.468 and 14.05 Gyr, respectively).

There are two equations (equations (1) and (2)) in two unknowns. Assuming that $P^{U/Th}$ did not vary from one star to another, which would otherwise undermine any attempt to calculate ages based on the U/Th ratio^{5,6,8,10–12}, and that GCE and LMHS record the same T_G (refs 2, 4), then it is possible to solve these equations. This problem can be reduced to determining the intersection between the two curves representative of equations (1) and (2) in T_G – $P^{U/Th}$ space. Using the two curves derived from LMHS and GCE (Fig. 2), it is easy to determine the uncertainty ellipsoid of the intersection and then compute the marginal probability densities for both $P^{U/Th}$ and T_G . The radiometric age of the Galaxy is $14.5^{+2.8}_{-2.2}$ Gyr and $P^{U/Th}$ is $0.571^{+0.037}_{-0.031}$, if a gaussian rate of infall is adopted. These results are not very sensitive to the chosen parameterization of the infall rate. Indeed, an exponential infall yields identical results for $P^{U/Th}$ and T_G within uncertainties ($15.1^{+2.8}_{-2.3}$ Gyr and $0.602^{+0.058}_{-0.043}$, respectively). It is often stated that the U/Th ratio measured in LMHS provides a model independent age of the Milky Way. There is in fact no such thing as a model independent age for LMHS, as nuclear physics models are used to predict $P^{U/Th}$ (refs 2, 3, 5–8) and observations of $R_{LMHS}^{U/Th}$ rely on modelling of stellar atmospheres^{10–12}.

The radiometric age of the Milky Way ($14.5^{+2.8}_{-2.2}$ Gyr) agrees with independent estimates based on colour–magnitude diagrams of globular clusters ($12.2^{+2.2}_{-1.2}$ Gyr, ref. 14) and cooling sequences of white dwarfs (12.1 ± 1.3 Gyr, ref. 15). This value also agrees with earlier estimates of the age of LMHS^{5–8,10–12}. GCE models had already been used to refine the value of the production ratio¹⁸ but the approach was circular, as an age was assumed to derive $P^{U/Th}$ using GCE, which was then used to choose acceptable nuclear physics models^{5,6}, which were used to calculate $P^{U/Th}$, which was eventually injected in LMHS to derive a new age^{5,6,11}. To summarize, an age was assumed in GCE to calculate an age using LMHS. The approach that is presented here avoids this circularity. Weighing the various estimates of T_G by their uncertainties, I obtain a more precise value of 12.5 ± 0.9 Gyr for the age of the Milky Way. For comparison, the age of the Universe is 13.7 ± 0.2 Gyr (ref. 13).

Determining $P^{U/Th}$ is crucial for chronometric applications. For this reason, a great deal of effort has been devoted over the last 50 yr to pin down this value precisely^{1–8}. If one assumes that all the progenitors of ^{232}Th and ^{238}U are produced in equal abundances on the r-process path, then determining $P^{U/Th}$ reduces to counting the progenitors of each isotope^{1,2}. This zeroth-order estimate gives a value of 0.578 for $P^{U/Th}$ (ref. 2). Interestingly, this turns out to be very close to the value that is obtained here, $0.571^{+0.037}_{-0.031}$. This similarity may be coincidental, as a compilation of $P^{U/Th}$ values between 1957 and 2003 shows that the estimates actually vary from 0.4 to 0.7 (Fig. 3). This question has received renewed attention in the past few years with the detections of U and Th in CS 31082–001 and BD +17°3248 (refs 10–12). Two groups tried to estimate $P^{U/Th}$ and assess its uncertainty in the framework of r-process calculations using state-of-the-art nuclear physics models. Goriely and Arnould⁵ estimated a value for $P^{U/Th}$ of $0.435^{+0.329}_{-0.137}$, while Schatz *et al.*⁶ proposed a more

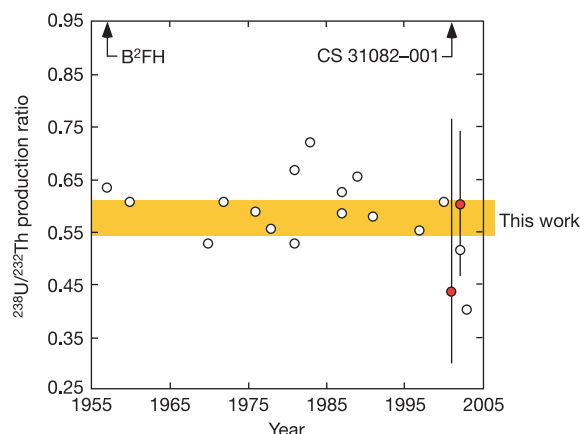


Figure 3 | Estimations of the $^{238}\text{U}/^{232}\text{Th}$ production ratio in r-process nucleosynthesis. The points are literature data (refs 1–8, see ref. 2 and references therein for publications before 1991) and the yellow band is the production ratio derived in this work ($0.571^{+0.037}_{-0.031}$, see text). The two arrows mark the first estimate of the U/Th ratio in the framework of r-process nucleosynthesis (labelled B^2FH , ref. 1) and the first determination of the U/Th ratio in a low metallicity star (CS 31082–001, ref. 10). The two red dots are state-of-the-art calculations aimed at estimating the U/Th production ratio and its possible uncertainty, arising primarily from uncertainties in the nuclear physics involved^{5,6}. As shown, GCE and LMHS provide tight constraints on the U/Th production ratio that can be used in turn to refine nuclear mass formulae^{5,6}, to help determine the source of Galactic cosmic rays, and to date circumstellar grains⁵.

restricted range of 0.603 ± 0.139 . Note that in the later study⁶, the authors had to rely on GCE calculations to rule out a nuclear mass model (HFBCS-1) that gave a lower $P^{U/Th}$ value of 0.455. The value that is derived here on the basis of GCE and LMHS ($0.571^{+0.037}_{-0.031}$) is much more precise than these two estimates (Fig. 3). The fact that the age of the Galaxy is consistent with results obtained using other methods^{13–15} provides confidence that the production ratio is also accurate. This is, to my knowledge, the first time that $P^{U/Th}$ has been calculated on the basis of GCE and the Solar System U/Th ratio without making any assumption about T_G , and the first time that a probabilistic meaning can be ascribed to the uncertainty quoted for the production ratio ($+0.037/-0.031$, 68% confidence interval). For comparison, $P^{U/Th}$ obtained from equation (2) alone and assuming that LMHS formed 0.5 Gyr after the Big Bang is 0.465 ± 0.117 , a value that is much less precise than the ratio obtained combining GCE and LMHS.

Because the r-process follows a path that is very distant from the valley of β -stability^{1–8}, on the neutron-rich side, the relevant nuclear properties used in the calculations cannot be measured and have to be inferred from extrapolations of theories that are anchored to measured nuclei^{5,6}. The calculations are in particular very sensitive to the formalism that is used to calculate nuclear masses^{5,6}, which directly affect the fission probabilities and the rates of neutron captures, photodisintegrations and β -decays. Goriely and Arnould⁵ and Schatz *et al.*⁶ tested various mass formulae in r-process calculations and used earlier work on GCE¹⁸ to rule out some of them. When the astrophysical conditions under which r-process nucleosynthesis occurs are clearly identified, the production ratio that is derived here could be used to constrain possible mass formulae and improve on r-process model predictions for other important cosmochronometers, namely ^{235}U , ^{244}Pu and ^{247}Cm . The U/Th production ratio is also crucial to help determine the source material of Galactic cosmic rays and possibly date circumstellar silicon carbide and diamond grains⁵. These grains are found in primitive meteorites and are thought to have condensed in the outflows of massive stars, before Solar System formation. With recent advances in mass spectrometers, U–Th dating of such material may become feasible in the near future.

Received 17 February; accepted 14 April 2005.

- Burbidge, E. M., Burbidge, G. R., Fowler, W. A. & Hoyle, F. Synthesis of the elements in stars. *Rev. Mod. Phys.* **29**, 547–650 (1957).
- Cowan, J. J., Thielemann, F.-K. & Truran, J. W. Radioactive dating of the elements. *Annu. Rev. Astron. Astrophys.* **29**, 447–497 (1991).
- Pfeiffer, B., Kratz, K.-L. & Thielemann, F.-K. Analysis of the solar-system r-process abundance pattern with the new ETFSI-Q mass formula. *Z. Phys. A* **357**, 235–238 (1997).
- Meyer, B. S. & Truran, J. W. Nucleocosmochronology. *Phys. Rep.* **333–334**, 1–11 (2000).
- Goriely, S. & Arnould, M. Actinides: How well do we know their stellar production? *Astron. Astrophys.* **379**, 1113–1122 (2001).
- Schatz, H. *et al.* Thorium and uranium chronometers applied to CS 31082–001. *Astrophys. J.* **579**, 626–638 (2002).
- Wanajo, S., Itoh, N., Ishimaru, Y., Nozawa, S. & Beers, T. C. The r-process in the neutrino winds of core-collapse supernovae and U–Th cosmochronology. *Astrophys. J.* **577**, 853–865 (2002).
- Otsuki, K., Mathews, G. J. & Kajino, T. r-Process abundance universality and actinide cosmochronology. *N. Astron.* **8**, 767–776 (2003).
- Chen, J. H., Wasserburg, G. J. & Papanastassiou, D. A. Th and U abundances in chondritic meteorites. *Lunar Planet. Sci. XXIV*, 277–278 (1993).
- Cayrel, R. *et al.* Measurement of stellar age from uranium decay. *Nature* **409**, 691–692 (2001).
- Hill, V. *et al.* First stars. I. The extreme r-element rich, iron-poor halo giant CS 31082–001. *Astron. Astrophys.* **387**, 560–579 (2002).
- Cowan, J. J. *et al.* The chemical composition and age of the metal-poor halo star BD +17°3248. *Astrophys. J.* **572**, 861–879 (2002).
- Spergel, D. N. *et al.* First-year Wilkinson Microwave Anisotropy Probe (WMAP) observations: determination of cosmological parameters. *Astrophys. J. Suppl. Ser.* **148**, 175–194 (2003).
- Krauss, L. M. & Chaboyer, B. Age estimates of globular clusters in the Milky Way: constraints on cosmology. *Science* **299**, 65–69 (2003).
- Hansen, B. *et al.* Hubble Space Telescope observations of the white dwarf cooling sequence of M4. *Astrophys. J. Suppl. Ser.* **155**, 551–576 (2004).
- Rutherford, E. Origin of actinium and the age of the Earth. *Nature* **123**, 313–314 (1929).
- Tinsley, B. M. Evolution of the stars and gas in galaxies. *Fund. Cosmic Phys.* **5**, 287–388 (1980).
- Yokoi, K., Takahashi, K. & Arnould, M. The ^{187}Re – ^{187}Os chronology and chemical evolution of the Galaxy. *Astron. Astrophys.* **117**, 65–82 (1983).
- Clayton, D. D. Nuclear cosmochronology within analytic models of the chemical evolution of the solar neighbourhood. *Mon. Not. R. Astron. Soc.* **234**, 1–36 (1988).
- Pagel, B. E. J. *Nucleosynthesis and Chemical Evolution of Galaxies* (Cambridge Univ. Press, Cambridge, 1997).
- Lodders, K. Solar system abundances and condensation temperatures of the elements. *Astrophys. J.* **591**, 1220–1247 (2003).
- Nordström, B. *et al.* The Geneva-Copenhagen survey of the solar neighbourhood. *Astron. Astrophys.* **418**, 989–1019 (2004).
- Wakker, B. P. *et al.* Accretion of low-metallicity gas by the Milky Way. *Nature* **402**, 388–390 (1999).
- Dauphas, N., Rauscher, T., Marty, B. & Reisberg, L. Short-lived p-nuclides in the early solar system and implications on the nucleosynthetic role of X-ray binaries. *Nucl. Phys. A* **719**, 287c–295c (2003).
- Kennicutt, R. C. Jr The global Schmidt law in star-forming galaxies. *Astrophys. J.* **498**, 541–552 (1998).
- Holmberg, J. & Flynn, C. The local density of matter mapped by Hipparcos. *Mon. Not. R. Astron. Soc.* **313**, 209–216 (2000).
- Holmberg, J. & Flynn, C. The local surface density of disc matter mapped by Hipparcos. *Mon. Not. R. Astron. Soc.* **352**, 440–446 (2004).
- Chang, R. X., Hou, J. L., Shu, C. G. & Fu, C. Q. Two-component model for the chemical evolution of the galactic disk. *Astron. Astrophys.* **350**, 38–48 (1999).
- Argast, D., Samland, M., Gerhard, O. E. & Thielemann, F. K. Metal-poor halo stars as tracers of ISM mixing processes during halo formation. *Astron. Astrophys.* **356**, 873–887 (2000).
- Chiappini, C., Matteucci, F., Beers, T. C. & Nomoto, K. The earliest phases of Galaxy formation. *Astrophys. J.* **515**, 226–238 (2000).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements I thank J. W. Truran, T. Rauscher, A. M. Davis, J. J. Cowan, R. Cayrel, V. Hill, R. Yokochi, B. Marty and L. Reisberg for discussions, and B. Hansen for communication of the 68% confidence interval for the age of the globular cluster Messier 4.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The author declares no competing financial interests. Correspondence and requests for materials should be addressed to N.D. (dauphas@uchicago.edu).

LETTERS

Neutron and X-ray diffraction study of the broken symmetry phase transition in solid deuterium

Igor Goncharenko¹ & Paul Loubeyre²

The solid hydrogen compounds D₂, HD and H₂ remain quantum molecular solids up to pressures in the 100 GPa range¹. A remarkable macroscopic consequence is the existence of a pressure-induced broken symmetry phase transition^{2–4}, in which the molecules go from a spherical rotational state to an anisotropic rotational state. Theoretical understanding of the broken symmetry phase structure remains controversial, despite numerous studies^{5–10}. Some open questions concern the existence of long- or short-range orientational order; whether a strong isotopic shift on the transition pressure should be assigned to the nuclear zero-point motion or to quantum localization; and whether the structures are cubic, hexagonal or orthorhombic. Here we present experimental data on the structure of the broken symmetry phase in solid D₂, obtained by a combination of neutron and X-ray diffraction up to 60 GPa. Our data are incompatible with orthorhombic structures predicted by recent theoretical works. We find that the broken symmetry phase structure is incommensurate with local orientational order, being similar to that found in metastable cubic *para*-D₂.

In the ground state at low temperature and low pressure, hydrogen molecules are in the $J = 0$ spherical rotational states¹¹. When pressure is applied, the molecules are driven closer to each other, so that at high enough pressure the new equilibrium state of the system is given by a trade-off between going higher in kinetic energy (that is, to $J \neq 0$ rotational levels) and gaining a negative potential energy through a orientational ordering that minimizes electric quadrupole–quadrupole (EQQ) energy. First-principles calculations show¹² that the EQQ interactions dominate over other interactions in the pressure range up to 100 GPa. Surprisingly, the increase in ordering by the admixture of the $J \neq 0$ states is not gradual. A first-order phase transition from the orientationally disordered phase to the orientationally ordered phase was first predicted¹³ in *para*-H₂ and *ortho*-D₂ at a pressure of ~15–30 GPa, accompanied by a structural transition from hexagonal close packed (h.c.p.) to face-centred cubic (f.c.c.). It was then observed that the broken symmetry phase (BSP) transition in fact occurs with a strong isotopic shift in the transition pressure, respectively 28 GPa for D₂ (ref. 2), 69 GPa for HD (ref. 4) and 110 GPa for H₂ (ref. 3). The orientationally disordered and ordered phases were also named as phase I and phase II, respectively. Furthermore, detailed spectroscopic studies have shown that the persistence of the optical phonon in phase II should imply that the h.c.p. structure of the molecular centres is not strongly affected by the orientational order¹⁴.

The calculation of the structure of the BSP poses a complex many-body problem, because it must embody the quantum character of the nuclei. A large number of calculations, based on different approximations, have tackled this problem, with some controversy. The first-principles calculations based on density functional theory (DFT) suggest *Cmc*2₁ (ref. 6), *Pca*2₁ (refs 7, 9, 10) or *P2*₁/c (ref. 8) structures in phase II. In these structures, the molecular centres do

not deviate significantly from the h.c.p. lattice, but the molecules are ordered in the a – b plane with a polar angle in the range 40–70°. All the above structures are orthorhombic, with four molecules per unit cell. Another approach is to use realistic interactions between the molecules, and to solve quasi-exactly the hamiltonian of the interacting nuclei by a path integral Monte Carlo simulation^{5,12}. One such calculation⁵ suggests a *Pa*3-type local orientation of the molecules and without displacement of their molecular centres from the h.c.p. lattice. No long-range structure was proposed in this work. Finally, it should be noted that the structure of phase II is an interesting solution of the model of coupled quantum rotors, which has been applied to various physical systems¹⁵.

The only direct experimental methods available to study crystal structure in the high-pressure phases of H₂ and D₂ are X-ray or neutron scattering experiments. D₂ or H₂ have no inner electronic shells, and X-rays are scattered by electrons in molecular orbitals. Therefore X-rays are almost insensitive to individual positions of H(D) atoms in the molecules and to orientational ordering. Contrarily, neutrons are scattered by individual nuclei, and therefore can be used to determine the individual positions of H(D) in the structure, as well as the orientations of the molecules. Because of the very small scattering power of a hydrogen crystal in a diamond anvil cell, the use of a third-generation synchrotron source and the growth of a single crystal in helium were previously needed to measure the equation of state of H₂ and D₂ to 120 GPa at 300 K (ref. 16). This X-ray approach has been extended here to low temperature. Owing to the low intensity of neutron sources, the high-pressure neutron study of the hydrogens is more challenging. Recent progress in pressure techniques and neutron instrumentation at the Laboratoire Léon Brillouin¹⁷ is used here to push the pressure limits for single-crystal neutron diffraction experiments. We present below a combination of neutron and X-ray techniques (up to 38 GPa and 60 GPa respectively, down to 1.5 K) to solve the structure of phase II.

Four different D₂ crystals were studied by X-ray diffraction: two of them were embedded in helium pressure transmitting medium. In one run, the single crystal of D₂ in helium was cooled at a constant pressure of 63.4 GPa and the (100), (101) and (002) reflections were measured. As seen in Fig. 1a, a small positive discontinuity in the lattice parameter c ($\Delta c/c = 3 \times 10^{-4}$) was observed at 70 K. A negative discontinuity in the parameter a ($\Delta a/a = -6 \times 10^{-4}$) was also observed at 70 K. This gives a transition point to phase II that is in excellent agreement with previous spectroscopic studies, and the corresponding volume discontinuity, $\Delta V/V = -10^{-3}$, confirms a previous estimate from the Clapeyron equation and the slope of the I–II boundary line^{1,18}. We note that the orthorhombic structures with the polar angle of the molecule almost in the a – b plane should give a negative discontinuity in c of about 1% (ref. 7), whereas a positive discontinuity is observed here. Two other runs have followed the evolution of reflections of the [100] and [101] class by going up in pressure at 25 K. One crystal had helium as pressure transmitting

¹Laboratoire Léon Brillouin CEA-CNRS, CEA Saclay, 91191 Gif-sur-Yvette, France. ²Département Physique Théorique et Applications, CEA, 91680 Bruyères-le-Châtel, France.

medium and the other did not. As seen in Fig. 1b, no discontinuity of the 100 *d*-spacing could be detected within the accuracy of the measurement. From X-ray measurements, we find that the I–II transition does not affect the diffraction peaks of the h.c.p. lattice, except for a very small volume discontinuity. Yet, as seen below, a transition was observed in both cases at around 16 GPa, as determined from the appearance of a superstructural peak.

One single crystal of D₂ in helium was devoted to the neutron measurements. The mosaicity and the orientation matrix of the single crystal were first determined by X-ray diffraction. The neutron data were collected at 38 GPa in the temperature range 1.5–70 K, above and below the I–II transition (I–II transition temperature $T_{\text{I-II}} \approx 44$ K at 38 GPa). The (100), (010), ($\bar{1}\bar{1}0$), (101), (110), ($\bar{1}\bar{1}0$) and (002) reflections were collected. As seen in Fig. 2, the quality of the crystal (rocking curve of 1°) allows an accurate determination of the intensity of the various reflections (apart from the (002) reflection, which was contaminated by a diamond reflection). In Fig. 2, integrated intensities of various neutron reflections are plotted versus temperature and compared to the expected behaviour of the strongest candidates for the structure of phase II. Surprisingly, none is satisfactory. All being of an orthorhombic symmetry, they give non-equivalent intensities for (100), (010) and ($\bar{1}\bar{1}0$) reflections (all indices are given in the hexagonal unit cell), whereas exactly the same intensity is observed (see inset in Fig. 2b). We could then assume that the orthorhombic symmetry is shadowed by twinning

along the crystallographic directions linked by the *P*3 symmetry of phase I. The sample would then consist of three equivalently populated orthorhombic domains, with *a* axes directed along the [100], [010] or [$\bar{1}\bar{1}0$] directions of the unit cell of phase I. Even under this assumption, the predicted intensities from *Cmca* and *Cmc*₂₁ structures completely disagree with the experimental data. *P*₂₁/*c* and *Pca*₂₁ structures explain the intensity of the (101) reflection, but not that of the (100) and (110) reflections. Assuming any non-equivalent domain distribution would only increase the difference between experiment and the predicted structures.

To explain the data, we follow the theoretical prediction⁵ that the I–II transition is purely rotational with a *Pa*3-type local order. The *Pa*3 structure is the structure observed in pure *ortho*-D₂ or *para*-H₂ (metastable in the *J* = 1 rotational state), which orders at low temperature in a cubic close-packed structure with neighbouring molecules directed along the different cubic diagonals¹¹. This structure minimizes the EQQ energy for a three-dimensional compact structure. To transpose a similar type of order on a hexagonal lattice, the molecules form four sublattices with molecular axes directed perpendicular to the four different faces of a tetrahedron. The

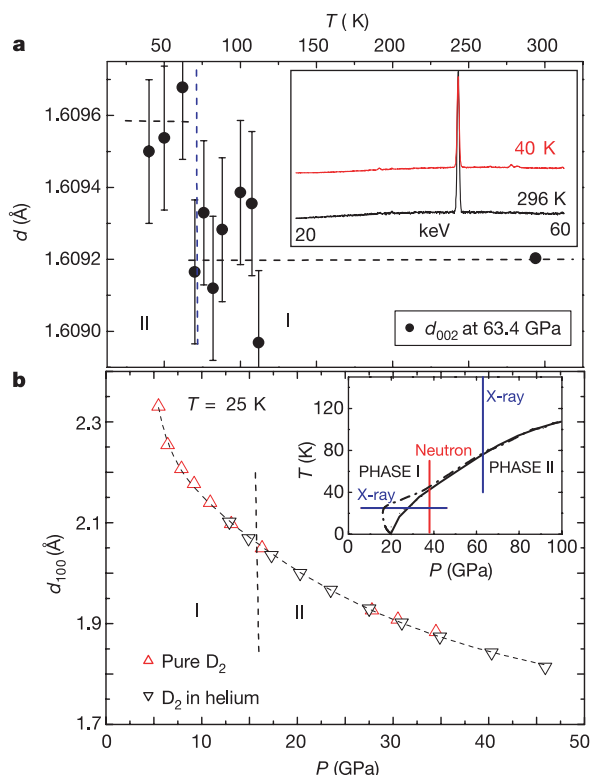


Figure 1 | X-ray diffraction measurements at the I–II phase transition in solid D₂. **a**, Evolution with temperature (*T*) of the *d*-spacing at 63.4 GPa, as measured by EDX. Filled circles are data points; error bars correspond to the 2×10^{-4} precision of the EDX measurement. The inset shows the (002) peak measured in phase I at 296 K (black) and in phase II at 40 K (red). **b**, Evolution with pressure (*P*) of the *d*-spacing of the (100) reflections measured by ADX at 25 K. The black and red triangles indicate data from a single crystal of D₂ respectively with and without helium pressure transmitting medium. The inset shows the pressure–temperature domains of the X-ray (in blue) and neutron (in red) measurements in the phase diagram. The solid and dashed black lines show respectively the I–II boundary line as determined from spectroscopic data^{14,18} and the re-entrant behaviour proposed here.

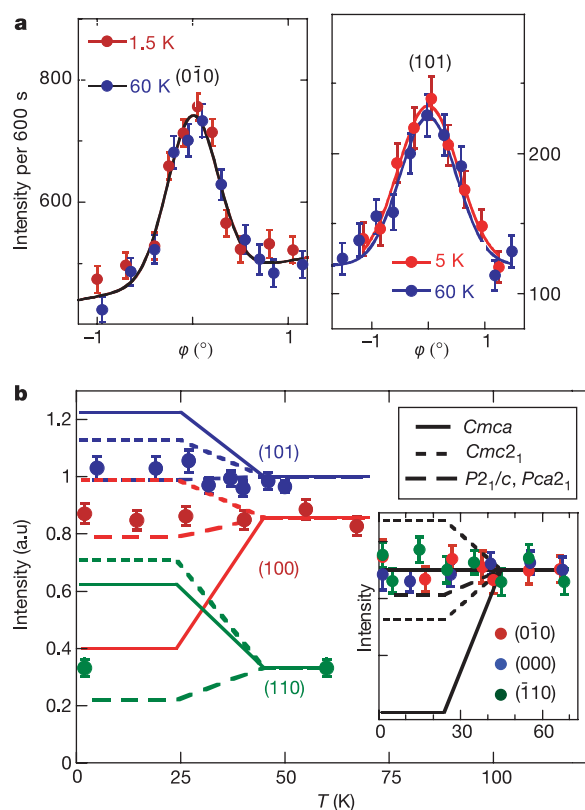


Figure 2 | Neutron diffraction measurements at the I–II phase transition on a single crystal of D₂ at 38 GPa with helium pressure transmitting medium. Error bars are statistical errors (data are shown as mean $\pm$ standard deviation) in the neutron experiment. Other contributions to the errors are negligible compared to the statistical error. **a**, Neutron intensities of the (010) and (101) reflections versus rocking angle ϕ in phase I (blue) and phase II (red). **b**, Comparison between experiment and the predicted neutron intensity for various structures. The predicted intensities have been averaged over the reflections linked by the *P*3 symmetry assuming an equivalent domain distribution of twin crystals. Also, the atomic displacements $\langle U^2 \rangle$ given in ref. 6 (scaled to a pressure of 38 GPa and for the mass of a deuteron) were included in the calculation. Filled circles, experiments; horizontal lines, calculated intensities for different orthorhombic models; sloping lines, guides for the eye. Inset, calculation for a single domain crystal of the (100), (010) and (110) reflections linked by *P*3 symmetry.

molecules are then arranged so that every molecule is surrounded by molecules belonging to the other three sublattices. A schematic drawing of the structure is shown in Fig. 3a (projection in the a - b plane). The unit cell is of $P\bar{3}$ symmetry, and consists of eight molecules. The neutron intensities calculated with this structure are seen in Fig. 3c to be in good agreement with the experimental data.

The $Pa\bar{3}$ -type of local order has a topological frustration when built in a hexagonal lattice. Whereas in the cubic structure the four sublattices correspond to four high-symmetry directions, in the h.c.p. structure the four orientations should be chosen from seven possible directions in the $5 \times D_2$ bipyramid (perpendicular to three 'up' and three 'down' faces and one along the c axis; Fig. 3b). There are no physical reasons to choose between 'up' and 'down' directions. An attempt to avoid the frustration by fixing the molecular axis in the a - b plane leads to the $P6_3/m$ structure previously proposed^{19,20} for $J = 1$ hydrogen. But, as shown in Fig. 3c, this model does not satisfy the neutron data. Stacking faults between the 'up' and 'down'

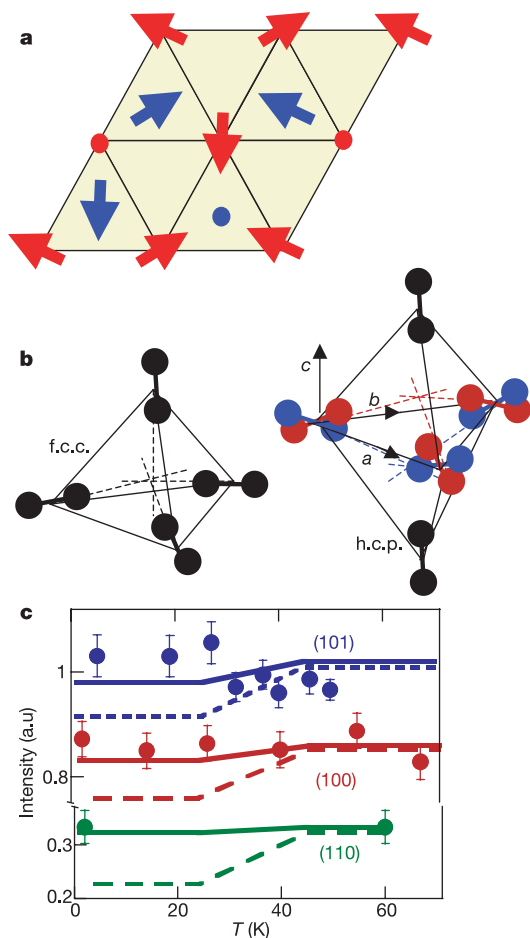


Figure 3 | Proposed structure for phase II of D_2 . **a**, Projection of the $P\bar{3}$ local structure to the a - b plane. Red and blue colours indicate molecules centred in the planes $z = 0$ and $z = 0.5$, respectively. Arrows point to the positive z -hemisphere, circles depict molecules lying along the c -axis. **b**, Left, orientational order in the f.c.c. lattice ($Pa\bar{3}$ structure). Molecular axes are directed towards the centre of the tetrahedron. Right, frustration in the h.c.p. lattice. In red and blue we show two different orientations (towards the centres of the upper and lower tetrahedrons, respectively) that have the same EQQ energy. Alternating 'red' and 'blue' configurations result in a superstructure in the $(a$ - b) plane. **c**, Calculated intensities for the $P\bar{3}$ (full lines) and $P6_3/m$ (dashed lines) structures versus experimental data (filled circles). Error bars are statistical errors (data are shown as mean $\pm$ standard deviation). Sloping lines are guides for the eye.

configurations should develop. If randomly distributed, they should give a short-range ordered structure. If collectively arranged, they should give an incommensurate long-range order. As shown in Fig. 4, incommensurate reflections with X-rays and with neutrons are indeed observed in phase II. The incommensurate vector is lying along the (100) direction. When the D_2 single crystal is embedded in helium, the incommensurate peak is observed along only one of the [100] directions, whereas when no pressure transmitting medium is used incommensurate peaks are observed along all three (100), (010) and (110) directions, because of twinning of the crystal under non-hydrostatic stress. No superstructure was observed along the c direction.

As seen in Fig. 4, the incommensurate peak appears reproducibly at the BSP transition, and is in fact the strongest structural signature of the phase transition. The transition is observed at 16 GPa at 25 K, whereas it has been measured between 20 GPa (ref. 14) and 24 GPa (ref. 18) at 5 K. This could be explained as follows: whereas at $T < 5$ K the equilibrium concentration of *para*- D_2 is practically zero ($< 0.1\%$), at 25 K it increases to $\sim 5\%$. The $J = 1$ states favour the BSP transition occurring at lower pressure, resulting in a re-entrant shape of the I-II boundary line (see Fig. 1b inset). The re-entrant behaviour has been clearly observed for HD (ref. 4) and was predicted for D_2 with equilibrium *ortho*-*para* concentration²¹. The incommensurate vector measured at 25 K by X-ray diffraction is essentially the same as the incommensurate vector measured by neutron diffraction at 1.5 K after the sample was kept for several days at this temperature in order to complete *para*-*ortho* conversion. Furthermore, we observed no changes in neutron intensities of the

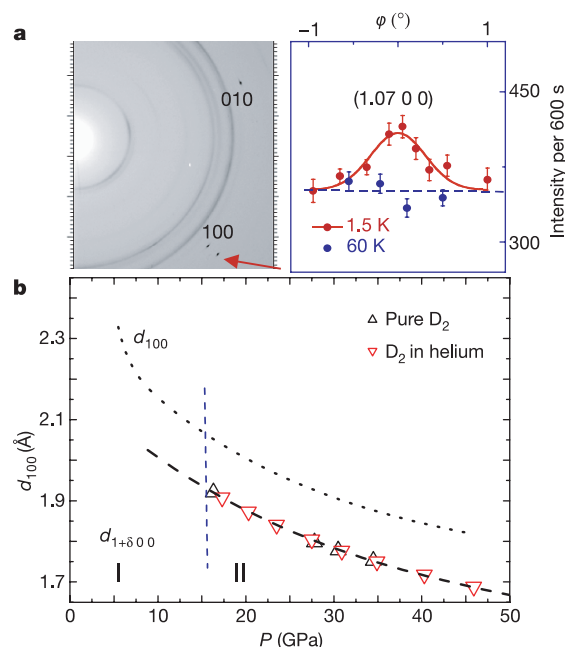


Figure 4 | Experimental evidence of a superstructure in phase II. **a**, Left, (100) and (010) reflections of a single crystal of D_2 in helium at 40 GPa. The diffraction peaks were recorded on an image plate detector during an oscillation of the diamond anvil cell (-20° to 20°) in the ADX configuration. The arrow indicates the (1.07 0 0) incommensurate peak. No incommensurate reflection is observed along the (010) reflection. The continuous diffraction rings are from the pressure cell. Right, neutron rocking curve of the incommensurate (1.07 0 0) reflection measured at 38 GPa. The incommensurate reflection reversibly disappears in phase I. The red and blue data points indicate respectively measurements in phase II and phase I. Error bars are statistical errors (data are shown as mean $\pm$ standard deviation). **b**, Evolution of d -spacing of the (1+ δ 0 0) peak versus pressure at 25 K. In D_2 with no pressure transmitting medium (black triangles), or D_2 in helium (red triangles), the incommensurate peak appears above 16 GPa.

structural peaks measured at 38 GPa between 25 K and 1.5 K and over a few days of equilibrium time. Hence, we believe that at 38 GPa the modest $J = 1$ concentrations have no drastic effect on the type of orientational ordering. The structure described above should be representative for pure *ortho*-deuterium. But additional measurements are required to investigate the influence of $J = 1$ states on specific regions of the pressure–temperature phase diagram, especially near the onset of the BSP transition.

The purely orientational transition on the h.c.p. lattice with *Pa3*-type local order seems to fit the structural data presented here. It also gives an explanation (involving frustration) of the incommensurate modulation of the structure. Also, the incommensurate modulation lowers the symmetry of the h.c.p. lattice, and hence could explain the multiplet structure of the roton bands and the additional vibron bands in the Raman and infrared spectra of phase II (ref. 14). Yet it remains to be understood why the *d*-spacing of the incommensurate peak is rather close to that of the (002) structural reflection. We note that the *Pa3*-type local order and incommensurate structures were not considered in DFT simulations. This should motivate further theoretical work. Finally, similar structural studies could be extended to higher pressures to characterize phase II of the other hydrogen isotopes, HD and H₂. It could be that the structure of phase II depends on the isotope²².

METHODS

A new hybrid diamond anvil cell has been developed to perform both X-ray (large axial X-ray opening) and neutron (large radial neutron opening) diffraction and to allow H₂/He gas loading. The X-ray measurements were performed by energy dispersive X-ray (EDX; 1 run) and angular dispersive X-ray (ADX; 3 runs) techniques at the ESRF (ID09 and ID 30 beamlines). The (002) reflection was accessible in the EDX run only. The neutron diffraction experiments were performed on diffractometer 6T2 at the Laboratoire Léon Brillouin. Pressure was measured by using standard ruby fluorescence techniques corrected for low temperatures. The *ortho*-*para* conversion rate is expected to increase under very high pressures¹⁸. The conversion rate under pressure has been measured for hydrogen^{23,24}, but no quantitative measurements for solid deuterium have been performed so far. The samples were allowed 12–72 h to convert to an equilibrium *ortho*-*para* ratio. This procedure is consistent with the earlier spectroscopic works¹⁸.

Received 21 January; accepted 28 April 2005.

- Mazin, I., Hemley, R., Goncharov, A., Hanfland, M. & Mao, H. K. Quantum and classical orientational ordering in solid hydrogen. *Phys. Rev. Lett.* **78**, 1066–1069 (1997).
- Silvera, I. F. & Wijngaarden, R. J. New low-temperature phase of molecular deuterium at ultrahigh pressure. *Phys. Rev. Lett.* **47**, 39–42 (1981).
- Lorenzana, H. E., Silvera, I. F. & Goettel, K. A. Evidence for a structural phase transition in solid hydrogen at megabar pressures. *Phys. Rev. Lett.* **63**, 2080–2083 (1989).
- Moshary, F., Chen, N. & Silvera, I. Remarkable high pressure phase line and orientational order in solid hydrogen deuteride. *Phys. Rev. Lett.* **71**, 3814–3817 (1993).
- Cui, T., Cheng, E., Alder, B. J. & Whaley, K. B. Rotational ordering in solid deuterium and hydrogen: A path integral Monte Carlo study. *Phys. Rev. B* **55**, 12253–12266 (1997).
- Kitamura, H., Tsuneyuki, Sh., Tadashi, O. & Miyake, T. Quantum distribution of protons in solid molecular hydrogen at megabar pressures. *Nature* **404**, 259–262 (2000).
- Kohanoff, J., Scandolo, S., Chiarotti, G. L. & Tosatti, E. Solid molecular hydrogen: The broken symmetry phase. *Phys. Rev. Lett.* **78**, 2783–2786 (1997).
- Johnson, K. A. & Ashcroft, N. W. Structure and bandgap closure in dense hydrogen. *Nature* **403**, 632–635 (2000).
- Stadele, M. & Martin, R. Metallization of molecular hydrogen: prediction from exact-exchange calculations. *Phys. Rev. Lett.* **84**, 6070–6073 (2000).
- Nagao, K., Takezawa, T. & Nagara, H. Ab initio calculation of optical-mode frequencies in compressed solid hydrogen. *Phys. Rev. B* **59**, 13741–13753 (1999).
- Van Kranendonk, J. Solid hydrogen: theory of the properties of solid H₂, HD, D₂ (Plenum, New York, 1983).
- Kaxiras, E. & Guo, Z. Orientational order in dense molecular hydrogen: A first-principles path-integral Monte Carlo calculation. *Phys. Rev. B* **49**, 11822–11832 (1994).
- Raich, J. & Eters, R. Rotational molecular motion in solid H₂ and D₂ under pressure. *J. Low Temp. Phys.* **6**, 229–240 (1972).
- Goncharov, A. F., Eggert, J. H., Mazin, I. I., Hemley, R. J. & Mao, H. K. Raman excitations and orientational ordering in deuterium at high pressure. *Phys. Rev. B* **54**, R15590–R15593 (1996).
- Freiman, Y., Tretyak, S., Antsygina, T. & Hemley, R. Novel phase behavior in quantum rotors. *J. Low Temp. Phys.* **133**, 251–260 (2003).
- Loubeyre, P. et al. X-ray diffraction and equation of state of hydrogen at megabar pressures. *Nature* **383**, 702–704 (1996).
- Goncharenko, I. N. Neutron diffraction experiments in diamond and sapphire pressure cells. *High Press. Res.* **24**, 193–204 (2004).
- Cui, L., Chen, H., Leon, S. & Silvera, I. Megabar pressure triple point in solid deuterium. *Phys. Rev. Lett.* **72**, 3048–3051 (1994).
- Miyagi, H. & Nakamura, T. Ground state energy difference of hcp and fcc ortho-hydrogens. *Prog. Theor. Phys.* **37**, 641–660 (1967).
- James, H. Orientational order in solid ortho-hydrogen. Hexagonal close-packed molecular lattice. *Phys. Rev.* **167**, 862–874 (1968).
- Hetényi, B., Scandolo, S. & Tosatti, E. Theoretical evidence for a reentrant phase diagram in ortho-*para* mixtures of solid H₂ at high pressure. *Phys. Rev. Lett.* **94**, 125503 (2005).
- Goncharov, A., Hemley, R., Mao, H. K. & Shu, J. New high-pressure excitations in parahydrogen. *Phys. Rev. Lett.* **80**, 101–104 (1998).
- Pravica, M. & Silvera, I. F. NMR study of ortho-*para* conversion at high pressure in hydrogen. *Phys. Rev. Lett.* **81**, 4180–4183 (1998).
- Eggert, J., Karmon, E., Hemley, R., Mao, H. K. & Goncharov, A. Pressure-enhanced ortho-*para* conversion in solid hydrogen up to 58 GPa. *Proc. Natl Acad. Sci. USA* **96**, 12269–12272 (1999).

Acknowledgements We thank A. Goukasov and O. Makarova for help in neutron experiments, and R. LeToullec, F. Occelli, M. Hanfland and M. Mezouar for help with the X-ray work.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to I.G. (gonch@llb.saclay.cea.fr) or P.L. (paul.loubeyre@cea.fr).

LETTERS

Soft X-ray microscopy at a spatial resolution better than 15 nm

Weilun Chao^{1,2}, Bruce D. Harteneck¹, J. Alexander Liddle¹, Erik H. Anderson¹ & David T. Attwood^{1,2}

Analytical tools that have spatial resolution at the nanometre scale are indispensable for the life and physical sciences. It is desirable that these tools also permit elemental and chemical identification on a scale of 10 nm or less, with large penetration depths. A variety of techniques^{1–7} in X-ray imaging are currently being developed that may provide these combined capabilities. Here we report the achievement of sub-15-nm spatial resolution with a soft X-ray microscope—and a clear path to below 10 nm—using an overlay technique for zone plate fabrication. The microscope covers a spectral range from a photon energy of 250 eV (~ 5 nm wavelength) to 1.8 keV (~ 0.7 nm), so that primary K and L atomic resonances of elements such as C, N, O, Al, Ti, Fe, Co and Ni can be probed. This X-ray microscopy technique is therefore suitable for a wide range of studies: biological imaging in the water window^{8,9}; studies of wet environmental samples^{10,11}; studies of magnetic nanostructures with both elemental and spin-orbit sensitivity^{12–14}; studies that require viewing through thin windows, coatings or substrates (such as buried electronic devices in a silicon chip¹⁵); and three-dimensional imaging of cryogenically fixed biological cells^{9,16}.

The microscope XM-1 at the Advanced Light Source (ALS) in Berkeley¹⁷ is schematically shown in Fig. 1. The microscope type is similar to that pioneered by the Göttingen/BESSY group (ref. 18, and references therein). A ‘micro’ zone plate (MZP) projects a full-field image to an X-ray-sensitive CCD (charge-coupled device), typically in one or a few seconds, often with several hundred images per day. The field of view is typically 10 μm , corresponding to a magnification of 2,500. The condenser zone plate (CZP), with a central stop, serves two purposes in that it provides partially coherent hollow-cone illumination², and, in combination with a pinhole, serves as the

monochromator. Monochromatic radiation of $\lambda/\Delta\lambda = 500$ is used. Both zone plates are fabricated in-house, using electron beam lithography¹⁹.

The spatial resolution of a zone plate based microscope is equal to $k_1\lambda/\text{NA}_{\text{MZP}}$ where λ is the wavelength, NA_{MZP} is the numerical aperture of the MZP, and k_1 is an illumination dependent constant, which ranges from 0.3 to 0.61. For a zone plate lens used at high magnification, $\text{NA}_{\text{MZP}} = \lambda/2\Delta r_{\text{MZP}}$ where Δr_{MZP} is the outermost (smallest) zone width of the MZP²⁰. For the partially coherent illumination^{21,22} used here, $k_1 \approx 0.4$ and thus the theoretical resolution is $0.8\Delta r_{\text{MZP}}$, as calculated using the SPLAT computer program²³ (a two-dimensional scalar diffraction code, which evaluates partially coherent imaging). In previous results with a $\Delta r_{\text{MZP}} = 25$ nm zone plate, we reported² an unambiguous spatial resolution of 20 nm. Here we describe the use of an overlay nanofabrication technique that allows us to fabricate zone plates with finer outer zone widths, to $\Delta r_{\text{MZP}} = 15$ nm, and to achieve a spatial resolution of below 15 nm, with clear potential for further extension.

This technique overcomes nanofabrication limits due to electron beam broadening in high feature density patterning. Beam broadening results from electron scattering within the recording medium (resist), leading to a loss of image contrast and thus resolvability for dense features. This effect is reduced by writing only semi-isolated features. Here the dense zone plate pattern is subdivided into two less dense, complementary patterns, which are fabricated separately and then overlaid with high accuracy to yield the desired pattern. The overlay technique allows us to achieve pattern densities several times higher than would otherwise be possible. The required placement accuracy for zone plates is typically one-third the smallest feature size—thus about 5 nm for the optic reported here with a 15 nm outer zone width. As described below, the zone placement accuracy achieved here is better than 2 nm across the two-dimensional field, leaving significant room for further zone plate advances. Note that this high placement accuracy overlay technique permits the achievement of smaller zones without incurring the low diffraction

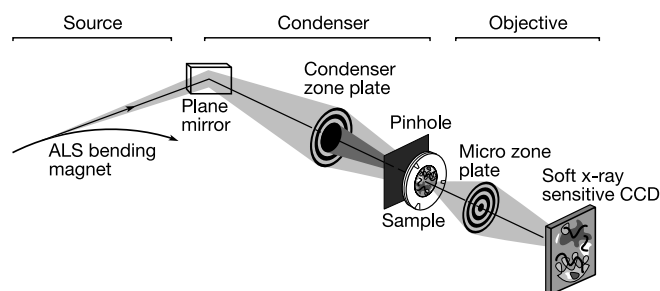


Figure 1 | A diagram of the soft X-ray microscope XM-1. The microscope uses a micro zone plate to project a full field image onto a CCD camera that is sensitive to soft X-rays. Partially coherent, hollow-cone illumination of the sample is provided by a condenser zone plate. A central stop and a pinhole provide monochromatization.



Figure 2 | An illustration of the overlay nanofabrication technique for micro zone plate fabrication. The zone plate is composed of even-numbered opaque zones (black and grey) and odd-numbered transparent zones (white). Set I (black), containing zones 2, 6, 10 ..., and its complement, set II (grey), are fabricated sequentially to form the desired overlaid micro zone plate.

¹Center for X-ray Optics, Lawrence Berkeley National Laboratory, 1 Cyclotron Road, MS 2-400, ²Department of Electrical Engineering and Computer Sciences, University of California, Berkeley, California 94720, USA.

efficiency inherent to photon sieves³, where pinhole distribution must meet the same tight placement constraints.

For the 15 nm zone plates reported here, the odd-numbered zones are transparent, while the even-numbered are opaque. The even-numbered zones are subdivided into two complementary sets: set I, containing zones 2, 6, 10..., and set II, containing zones 4, 8, 12..., as illustrated in Fig. 2. The nanofabrication process will be described in further detail elsewhere (W.C., J.A.L., E.A. and B.H., manuscript in preparation). The complete zone plate fabrication was conducted in-house, using our vector-scan electron beam lithography tool, the Nanowriter¹⁹, which has a 6.5 nm diameter (full-width at half-maximum) 100 keV electron beam, and excellent zone placement accuracy. The latter property is achieved by a combination of nanometre beam position control, accurate distortion calibration, and the use of a pattern generator specially designed for curved shapes²⁴. The resist, polymethyl methacrylate (PMMA), was used for pattern recording, while all opaque zones were formed by gold plating in a multistep process¹⁹ for reasonable zone plate efficiency, in the operating spectral range from 250 eV to 1.8 keV.

High alignment accuracy is the key to the success of our technique. Before exposure of the zone plate patterns, the Nanowriter's beam deflection was calibrated, using an in-house alignment algorithm²⁵, to the pre-fabricated alignment marks on the resist-coated wafer. This technique greatly reduces systematic zone placement errors, allowing us to consistently achieve a subpixel placement accuracy of 1.7 nm (1 s.d.). In Fig. 3, a scanning electron micrograph shows the outer zone region of a 15 nm zone plate, revealing near perfect alignment of the opaque zones. The gold zones contain small gaps at various positions, and have widths larger than desired, thus reducing efficiency in these early results. We expect to improve these in the future. The gold plated zones are 80 nm thick, as needed for opacity and thus efficiency, giving an aspect ratio of 5:1. The calculated diffraction efficiency to first order is 6%. Transmission of the substrate and plating base, which support the zone plate, is 70%, so the expected zone plate efficiency is 4%. This is typical of early state-of-the-art zone plates, and is consistent with our observed exposure time.

Using this zone plate, the microscope was used to image several patterns, including 15.1 nm and 19.5 nm half-period test objects at 1.52 nm wavelength ($h\nu = 815$ eV), with a magnification of 7,600. For optimal signal to noise ratio, the exposure time was 62 s, with 10^4 counts per pixel in the $2,048 \times 2,048$ pixel array CCD detector. This exposure time is about 30 times longer than would be used in

normal imaging. In order to minimize the effect of photon and electronic noise and CCD pixel size on the image resolution, we chose to collect about 8 times more photons per pixel than is typical, for improved statistics (10^4 compared to 10^3 photons per pixel), and we used 4 times more pixels than is typical for high spatial resolution recording at the same magnification. That is, we used a $2,048 \times 2,048$ pixel CCD (1.6 nm $\times$ 1.6 nm equivalent pixel size) rather than a $1,024 \times 1,024$ CCD (3.2 nm $\times$ 3.2 nm effective pixel size). For the imaging experiments here, the microscope configuration (Fig. 1) used is as follows. MZP: $\Delta r_{MZP} = 15$ nm, 500 zones, 30 μ m diameter; CZP: $\Delta r_{CZP} = 40$ nm, 56,250 zones, 9.0 mm outer diameter, 5 mm diameter central stop; pinhole, 14 μ m. The degree of partial coherence^{21,22}, σ , is 0.38. The test objects used for these resolution studies were Cr/Si multilayer coatings²⁰ in cross-section; they were fabricated in-house, using magnetron sputtering and conventional transmission electron microscopy sample preparation techniques²⁶.

Images obtained with the two zone plate lenses, having outer zone widths of 25 nm and 15 nm, are shown in Fig. 4. The two images on the left, Fig. 4a and c, were obtained with the 25 nm zone plate, at a wavelength of 2.07 nm (600 eV). This photon energy is just above the

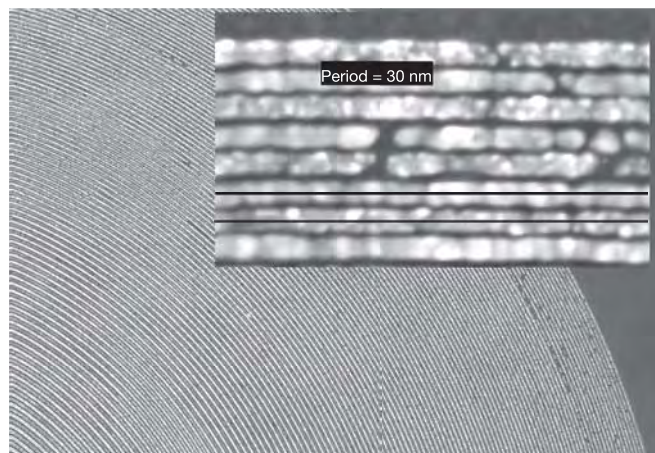


Figure 3 | Scanning electron micrograph of a zone plate with 15 nm outermost zone. Shown in the inset is a more detailed view of the outermost zones. The zonal period, as indicated by the two black lines, is measured to be 30 nm. The zone placement accuracy is measured to be 1.7 nm.

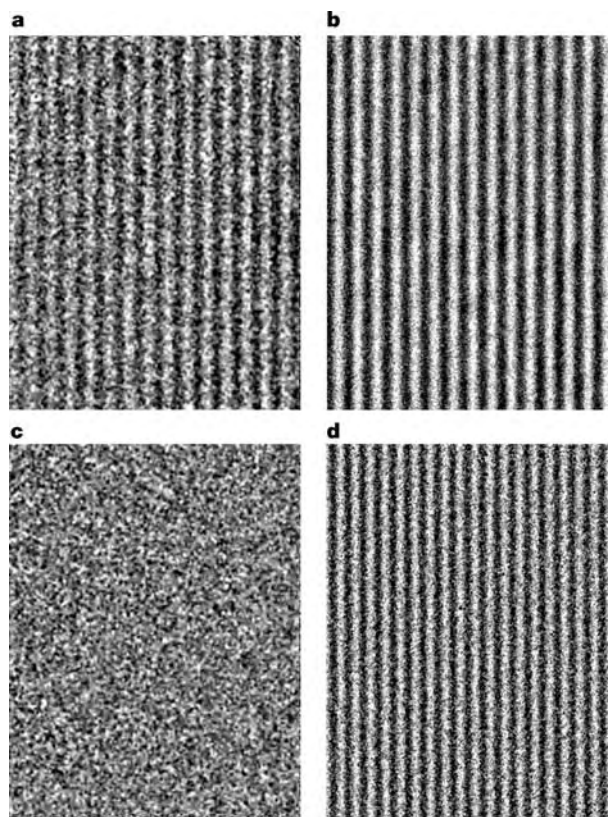


Figure 4 | Soft X-ray images of 15.1 nm and 19.5 nm half-period test objects, as formed with zone plates having outer zone widths of 25 nm and 15 nm. The test objects consist of Cr/Si multilayers, with 15.1 nm and 19.5 nm half-periods, respectively. Significant improvements are noted between the images obtained with the new 15 nm zone plate, as compared to earlier results obtained with the 25 nm zone plate. This is particularly evident for the 15 nm half-period images, for which the earlier result shows no modulation, whereas the image obtained with the 15 nm zone plate shows excellent modulation. **a**, Image of 19.5 nm half-period test object obtained previously with a 25 nm zone plate. **b**, Image of 19.5 nm half-period object with the 15 nm zone plate. **c**, Image of 15.1 nm half-period with the previous 25 nm zone plate. **d**, Image of 15.1 nm half-period with the 15 nm zone plate. Images **a** and **c** were obtained at a wavelength of 2.07 nm (600 eV photon energy); **b** and **d** were obtained at a wavelength of 1.52 nm (815 eV). The equivalent object plane pixel size for images **a** and **c** is 4.3 nm; the size for **b** and **d** is 1.6 nm.

Cr absorption edge at 574 eV. The image of the pattern with 19.5 nm lines and spaces (Fig. 4a) shows good modulation (20%), whereas the image of 15.1 nm lines and spaces (Fig. 4c) shows no modulation with this lens. As seen in Fig. 5, the latter data point is beyond cut-off for the 25 nm lens. Returning to Fig. 4, the two images on the right side, Fig. 4b and d, were obtained with the 15 nm zone plate lens, at a wavelength of 1.52 nm (815 eV). The shorter wavelength allowed us to maintain a convenient working distance. Images obtained with the 15 nm outer zone width lens (Fig. 4b and d) show clear improvements when compared to those with the 25 nm zone plate. The 19.5 nm image in Fig. 4b displays less noise and better contrast than that in Fig. 4a. The improvement is particularly evident in images of the 15.1 nm lines, for which the earlier results with the 25 nm lens showed no modulation (Fig. 4c), whereas the image obtained with the 15 nm lens (Fig. 4d) shows excellent modulation.

An intensity profile ('lineout') of Fig. 4d shows a modulation of 3,000 counts per pixel and a peak level of 40,000 counts per pixel in the CCD image, corresponding to a modulation of approximately 8%. To determine the optical modulation transfer function (MTF) of the system, knowledge of the intrinsic sample contrast is needed. This can be calculated using the known absorption parameters for Si and Cr, and the sample thickness. The sample thickness varies from zero (clear area) to many micrometres. The image was taken near the clear area, where we estimate the sample thickness to be approximately 50 nm (about a 3:1 aspect ratio, which generally provides good images). In this case, at a photon energy of 815 eV, the Si transmits 97% of the photons and the Cr transmits 64%. The image modulation, calculated using SPLAT for an ideal sample with equal 15 nm lines and spaces, is 20% for a perfect lens. Allowing for an imperfect lens with an MTF of 50%, close to its resolution limit, we expect only a 10% CCD image modulation. This value is within the error bars of the observed value. The same analysis applies to the 19.5 nm image in Fig. 4b, where the recorded CCD modulation is about 4,000 photons per pixel on a 36,000 photons per pixel background, or 11%. Unfortunately, because of the uncertainty in the sample thicknesses, and to a lesser extent the presence of somewhat non-uniform stray light, which affects the determination of image modulation, it is not

possible unambiguously to assign MTF values on the basis of these two measurements. It is worth noting, however, that modest image modulations with thin, high resolution samples are not unusual, and indeed those observed here are similar to those seen with our previously described 25 nm zone plate, which has been used successfully in many scientific studies.

Modelling of the MTF for the two lenses is shown by solid lines in Fig. 5. This computational modelling accounts for the partially coherent, hollow-cone, soft X-ray illumination employed in each case. With the higher numerical aperture ($NA = \lambda/2\Delta r$), the simulation predicts a proportionately increased resolving power, as indicated by the MTF shift to higher spatial frequencies. With this degree of partially coherent illumination, we calculate a theoretically achievable resolution of approximately $0.8\Delta r$, or 20 nm with the 25 nm zone plate and 12 nm with the 15 nm zone plate. This is consistent with the images in Fig. 4, and with the three data points in Fig. 5 for the 25 nm lens. Data points for the 15 nm lens (Fig. 4b and d) are not shown in Fig. 5 owing to insufficient knowledge of the sample thickness and the presence of somewhat non-uniform stray light, as mentioned above. We believe that the improved imaging capability, as illustrated in Fig. 4, and supported by Fig. 5, are the clearest demonstration of a significant advance with the present zone plate fabrication technique. In the near future, with the high electron beam placement accuracy and the ability of PMMA to support isolated features as small as 5 nm (ref. 27), we expect our overlay nanofabrication technique to yield high quality zone plates with outer zone widths of 10 nm, permitting a spatial resolution of 8 nm. Additional benefits of this overlay technique will be to permit higher aspect ratio zones for improved efficiency and, separately, the stacking of subzonal structures for additional improvement of efficiency, perhaps in a trade-off for resolution in multilevel zone plates²⁸.

In addition to improved spatial resolution, which scales as Δr , the zone plates reported here significantly reduce the depth of field, which scales²⁰ as $(\Delta r)^2$, thus offering a new capability for soft X-ray optical sectioning, and the further potential for improved spatial resolution soft X-ray tomography. This will require further efforts to model soft X-ray propagation through sequential two-dimensional image sections. In the life sciences, when combined with protein specific labelling⁸, these advances would permit quantitative protein localization in three-dimensional images of the cell, and thus permit studies of gene expression as a function of mutations, knockout genes, and so on (C. A. Larabell, personal communication).

With the existence of more than 30 synchrotron facilities worldwide, these advances in soft X-ray microscopy could be readily available to the research community. Furthermore, we anticipate that compact soft X-ray sources will also be available in the not too distant future, using laser-produced plasmas²⁹, femtosecond laser high harmonic techniques³⁰, or extreme-ultraviolet/soft X-ray lasers³¹. With these advances, we anticipate a wider use of zone plate based soft X-ray microscopy across the broad range of nanoscience and nanotechnology.

Received 3 September 2004; accepted 3 May 2005.

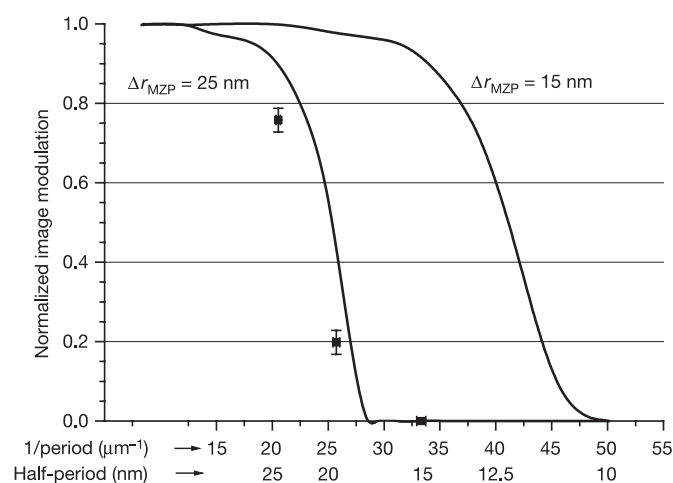


Figure 5 | The calculated modulation transfer functions of the microscope with two different zone plates. One zone plate has an outer zone width, Δr_{MZP} of 25 nm (left line), and the other has Δr_{MZP} = 15 nm (right line). The theoretical resolution for the two lenses are 19 nm and 12 nm, respectively. Also shown are data indicating the degree of modulation obtained for various test patterns using the Δr_{MZP} = 25 nm (squares). The Δr_{MZP} = 25 nm zone plate yielded 75% modulation for a half-period of 24.3 nm, 20% for 19.5 nm, and 0% modulation for a half-period of 15.1 nm (Fig. 4a and c). Values are means $\pm$ s.d. Using the Δr_{MZP} = 15 nm zone plate, image quality is much improved, as seen in Fig. 4b and d, but owing to uncertain sample thicknesses and stray light an accurate determination of the modulation was not possible.

1. Susini, J. & Joyeux, D. & Polack, F. (eds) *X-Ray Microscopy VII* (EDP Sciences, Paris, 2003).
2. Chao, W. *et al.* 20-nm-resolution soft x-ray microscopy demonstrated by use of multilayer test structures. *Opt. Lett.* **28**, 2019–2021 (2003).
3. Kipp, L. *et al.* Sharper images by focusing soft X-rays with photon sieves. *Nature* **414**, 184–188 (2001).
4. Eisebitt, S. *et al.* Lensless imaging of magnetic nanostructure by x-ray spectral-holography. *Nature* **432**, 885–888 (2004).
5. Miao, J., Charalambous, P., Kirz, J. & Sayre, D. Extending the methodology of X-ray crystallography to allow imaging of micrometre-sized non-crystalline specimens. *Nature* **400**, 342–344 (1999).
6. Miao, J. W. *et al.* Imaging whole *Escherichia coli* bacteria by using single-particle x-ray diffraction. *Proc. Natl Acad. Sci. USA* **100**, 110–112 (2003).
7. Marchesini, S. *et al.* X-ray image reconstruction from a diffraction pattern alone. *Phys. Rev. B* **68**, 140101–140104 (2003).

8. Meyer-Ilse, W. *et al.* High resolution protein localization using soft x-ray microscopy. *J. Microsc.* **201**, 395–403 (2001).
9. Larabell, C. A. & Le Gros, M. A. X-ray tomography generates 3-D reconstructions of the yeast, *Saccharomyces cerevisiae*, at 60-nm resolution. *Mol. Biol. Cell* **15**, 957–962 (2003); movie at (<http://ncxt.LBL.gov/movies/video-seg8.mov>).
10. Myneni, S. C. B., Brown, J. T., Martinez, G. A. & Meyer-Ilse, W. Imaging of humic substance macromolecular structures in water and soils. *Science* **286**, 1335–1337 (1999).
11. Juenger, M. C. G., Lamour, V. H. R., Monteiro, P. J. M., Gartner, E. M. & Denbeaux, G. P. Direct observation of cement hydration by soft X-ray transmission microscopy. *J. Mater. Sci. Lett.* **22**, 1335–1337 (2003).
12. Fischer, P., Schutz, G., Schmahl, G., Guttman, P. & Raasch, D. Imaging of magnetic domains with the X-ray microscope at BESSY using X-ray magnetic circular dichroism. *Z. Phys. B* **101**, 313–316 (1996).
13. Fischer, P. *et al.* Study of magnetic domains by magnetic soft x-ray transmission microscopy. *J. Phys. D* **35**, 2391–2397 (2002).
14. Stoll, H. *et al.* High-resolution imaging of fast magnetization dynamics in magnetic nanostructures. *Appl. Phys. Lett.* **84**, 3328–3330 (2004).
15. Schneider, G. *et al.* Electromigration in passivated Cu interconnects studied by transmission x-ray microscopy. *J. Vac. Sci. Technol. B* **20**, 3089–3094 (2002).
16. Schneider, G. *et al.* Computed tomography of cryogenic cells. *Surf. Rev. Lett.* **9**, 177–183 (2002).
17. Meyer-Ilse, W. *et al.* New high-resolution zone-plate microscope at Beamline 6.1 of the Advanced Light Source. *Synchr. Radiat. News* **8**, 29–33 (1995).
18. Schmahl, G. & Rudolph, D. (eds) *X-ray Microscopy* (Springer, Berlin, 1984).
19. Anderson, E. H. *et al.* Nanofabrication and diffractive optics for high-resolution x-ray applications. *J. Vac. Sci. Technol. B* **18**, 2970–2975 (2000).
20. Attwood, D. T. *Soft X-Rays and Extreme Ultraviolet Radiation* (Cambridge Univ. Press, Cambridge, UK, 2000).
21. Goodman, J. W. *Statistical Optics* 303–324 (Wiley, New York, 2000).
22. Born, M. & Wolf, E. *Principles of Optics* 7th edn, 441, 596–606 (Cambridge Univ. Press, Cambridge, UK, 1999).
23. Toh, K. K. H. & Neureuther, A. R. Identifying and monitoring effects of lens aberrations in projection printing. *Proc. SPIE* **772**, 202–209 (1987).
24. Anderson, E. H., Boegli, V. & Muray, L. P. Electron beam lithography digital pattern generator and electronics for generalized curvilinear structures. *J. Vac. Sci. Technol. B* **13**, 2529–2534 (1995).
25. Anderson, E. H., Ha, D. & Liddle, J. A. Sub-pixel alignment for direct-write electron beam lithography. *Microelectron. Eng.* **73–74**, 74–79 (2004).
26. Chao, W. *et al.* Demonstration of 20 nm half-pitch spatial resolution with soft x-ray microscopy. *J. Vac. Sci. Technol. B* **21**, 3108–3111 (2003).
27. Yasin, S., Hasko, D. G. & Ahmed, H. Fabrication of <5 nm width lines in poly(methylmethacrylate) resist using a water:isopropyl alcohol developer and ultrasonically-assisted development. *Appl. Phys. Lett.* **78**, 2760–2762 (2001).
28. Di Fabrizio, E. *et al.* High-efficiency multilevel zone plates for keV X-rays. *Nature* **401**, 895–898 (1999).
29. Vogt, U. *et al.* High-resolution spatial characterization of laser produced plasmas at soft x-ray wavelengths. *Appl. Phys. B* **78**, 53–58 (2004).
30. Gibson, E. A. *et al.* Coherent soft x-ray generation in the water window with quasi-phase matching. *Science* **302**, 95–98 (2003).
31. Larotonda, M. A. *et al.* Characteristics of a saturated 18.9 nm table top laser operating at 5 Hz repetition rate. *IEEE J. Select. Topics Quant. Electron.* **10**, 1363–1367 (2004).

Acknowledgements The authors acknowledge financial support from the National Science Foundation's Engineering Research Centre Program, the Department of Energy's Office of Science, Office of Basic Energy Sciences, and the Defense Advanced Research Projects Agency.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to W.C. (wlchao@lbl.gov).

LETTERS

Tomographic imaging using the nonlinear response of magnetic particles

Bernhard Gleich^{1*} & Jürgen Weizenecker^{1*}

The use of contrast agents and tracers in medical imaging has a long history^{1–7}. They provide important information for diagnosis and therapy, but for some desired applications, a higher resolution is required than can be obtained using the currently available medical imaging techniques. Consider, for example, the use of magnetic tracers in magnetic resonance imaging: detection thresholds for *in vitro*⁸ and *in vivo*⁹ imaging are such that the background signal from the host tissue is a crucial limiting factor. A sensitive method for detecting the magnetic particles directly is to measure their magnetic fields using relaxometry¹⁰; but this approach has the drawback that the inverse problem (associated with transforming the data into a spatial image) is ill posed and therefore yields low spatial resolution. Here we present a method for obtaining a high-resolution image of such tracers that takes advantage of the nonlinear magnetization curve of small magnetic particles. Initial ‘phantom’ experiments are reported that demonstrate the feasibility of the imaging method. The resolution that we achieve is already well below 1 mm. We evaluate the prospects for further improvement, and show that the method has the potential to be developed into an imaging method characterized by both high spatial resolution as well as high sensitivity.

We first introduce the general concept of magnetic particle imaging (MPI). MPI relies on the nonlinearity of the magnetization curves of ferromagnetic material and the fact that the particle magnetization saturates at some magnetic field strength. If an oscillating magnetic field, called the ‘modulation field’, is applied, with frequency f_1 and sufficiently high amplitude A , the magnetic material will exhibit a magnetization $M(t)$, where t is time. $M(t)$ contains not only the drive frequency f_1 , but also a series of harmonic frequencies (Fig. 1a). These higher frequencies can be easily separated from the received signal by means of appropriate filtering.

If the magnetic particles are also exposed to a time constant magnetic field with a sufficiently large magnitude, they saturate and the generation of harmonics is suppressed (Fig. 1b). Selective suppression of the harmonics is employed for the spatial encoding as follows. In addition to the modulation field, a time-independent field is superimposed (field plot in Fig. 2a) that vanishes in the centre of the imaging device (the field-free point, FFP) and increases in magnitude towards the edges. This field is called the ‘selection field’. If there is any magnetic material at the position of the FFP it will produce a signal containing higher harmonics. But only the magnetic material located at the FFP will respond to the a.c. modulation field. All other magnetic material remains in the state of saturation. By steering the FFP through the volume of interest, a tomographic image can be generated. (As a spatial variation in one direction of one field component is in general accompanied by a spatial variation of another component in another direction, only a single selection field is needed in MPI to obtain a three-dimensional spatial encoding.) The movement can be performed by moving the

whole coil assembly or by moving the object within the coil assembly. For simplicity, we assumed a low amplitude of the modulation field. Otherwise, the a.c. field shifts the FFP significantly.

In summary, to form an image, magnetic tracer material has to be applied to, or introduced into, the object. The object is placed in the selection field, and a weak magnetic modulation field is

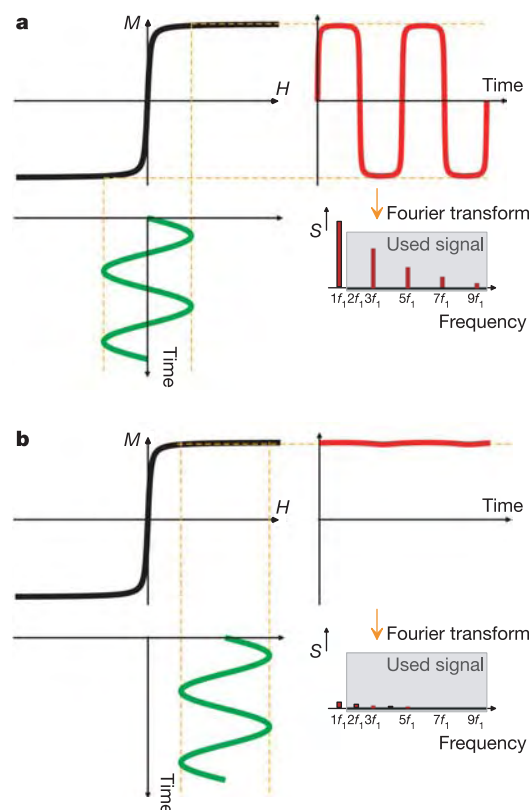


Figure 1 | Response of magnetic particles to an external magnetic field.

a, An oscillating magnetic field (H , modulation field, green curve) is applied to the magnetic material at a single frequency f_1 . As the magnetization curve (M , black curve) is nonlinear, the resulting time-dependent magnetization (red curve) exhibits higher harmonics, as is shown in the Fourier-transformed signal (S , red bars). **b**, A time-independent field is added to the modulation field. The oscillating field does not significantly change the magnetization of the material, as it is always in saturation. In this state, harmonics of the oscillating field are almost non-existent. The grey box indicates those harmonics used for image formation. The signal at f_1 is not used, as it is small compared to the superimposed induced modulation field signal, and therefore difficult to isolate.

¹Philips Research Hamburg, Röntgenstrasse 24-26, D-22335 Hamburg, Germany.

*These authors contributed equally to this work.

superimposed. Finally the object is moved (spatial encoding) to discrete positions and the magnitudes of the harmonics are recorded. An image of the magnetic tracer in the object is directly obtained by mapping the magnitude of the harmonics.

The method described so far is capable of generating images of the spatial distribution of magnetic material. However, the mechanical movement leads to low scanning speed and the signal to noise ratio (SNR) is low owing to the weak modulation field. The mechanical movement is dispensable if three additional orthogonal homogenous magnetic fields, called drive fields, are provided (Fig. 2b). The three components of the selection field can be cancelled, at any given point in space, by appropriate adjustment of these three fields. By driving each coil pair with a predefined current waveform, the FFP can be moved on a continuous trajectory over the object. This set-up is analogous to a mechanical motion set-up.

By using drive fields, it is possible to accelerate the movement of the FFP dramatically. For this purpose, a different sinusoidal current with a high frequency is applied to each coil pair. The amplitudes of the currents must be large enough to generate magnetic fields capable of cancelling the selection field at the border of the desired region of interest. The fast FFP movement leads to a rapid local change in magnetization as soon as the FFP passes a location containing magnetic material. The magnetization change induces a signal in the recording coil that exhibits higher harmonics of the drive field frequencies. This induced signal is sufficient for image reconstruction. The modulation field with low amplitude (Fig. 1) is now obsolete. Consequently, the introduction of the drive fields overcomes both drawbacks mentioned above, namely the low encoding speed and the low SNR.

So the spatial encoding can be realized in two ways, as follows: (1) using mechanical movement, or (2) using field-induced movement of the FFP. Furthermore, both possibilities can be combined. This is the situation realized in the present experimental set-up.

The two-dimensional object used for imaging consisted of distinct holes filled with an undiluted ($0.5 \text{ mol Fe l}^{-1}$), commercially available contrast agent (Resovist¹¹, Schering AG Berlin). Components used in the present set-up are illustrated in Fig. 2a. In order to form two-dimensional images, the object can be moved in two dimensions using a robot. Additionally, the drive field moves the FFP in the

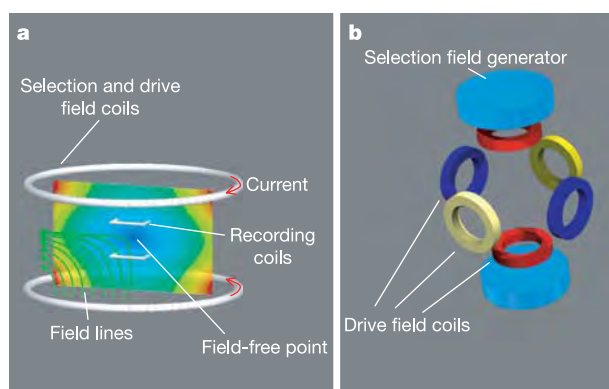


Figure 2 | The main components of the experiment, and an MPI scanner concept. **a**, The two large rings generate the selection field. Hence, a d.c. current with opposite direction in the upper and lower coil produces the sketched field (field lines and colour coded field magnitude) with the field-free point (FFP) in the centre. The same two rings serve as drive field coils, as an a.c. current is superimposed on the d.c. current. A pair of quadratic recording coils in the centre records the generated a.c. response (harmonics). **b**, The field-generating components are sketched schematically for an MPI scanner capable of encoding purely by drive fields. Two field generators produce the selection field. For each direction in space, two opposing drive field coils are used. These coils produce a more or less homogeneous field in the centre of the scanner and can therefore move the FFP.

vertical direction. Two alternative encoding types are demonstrated. First, the robot alone is used for the spatial encoding (mechanical FFP movement). However, a drive field ($A = 10 \text{ mT } \mu_0^{-1}$) for the generation of the harmonics is used in place of the low-amplitude modulation field, owing to the higher achievable SNR. Second, the robot provides the spatial encoding in the horizontal direction, while the drive field moves the FFP in the vertical direction. The same experiment was evaluated in both cases, but in the second case only a subset of robot scan positions was used.

Figure 3a shows an image of the object, for the case of pure mechanical movement of the object. Using a drive field leads to a contribution of neighbouring points to the recorded signal at a given robot position. This means that the simple method for generating an image (for the case of the weak modulation field) by mapping the magnitude of the harmonics is not appropriate, and a reconstruction is necessary; see Methods section.

In Fig. 3b the drive field encodes vertically and the robot is used for the horizontal encoding. As the drive field amplitude was not sufficiently high to move the FFP over the whole object, three separate images were reconstructed. They covered the upper, middle and lower regions of the object, and were averaged to form Fig. 3b. As a result, the resolution in the vertical direction is better than 0.3 mm . In the horizontal direction, the resolution is about 0.5 mm , as the derivative of the field component of the selection field in that direction is lower. The resolution in both cases (Fig. 3a and b) is the same, although the SNR differs owing to the shorter measurement time in the second case.

The theoretically expected resolution (R) is given by the ratio

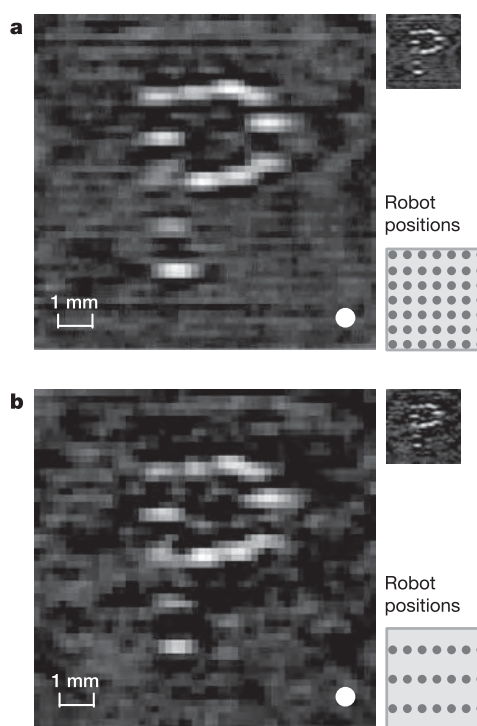


Figure 3 | Reconstructed images of the object for two different encoding types. The true size of the holes is indicated in the lower right corner of the large images. The drawings on the right side sketch the robot positions used for measurement (bottom) and a true scale image (top). In **a**, the data at all 52×52 robot positions were used, whereas in **b** only the data of 3×52 robot positions contribute to the reconstruction. In **a**, encoding is purely done by robot movement, although the FFP moves a considerable distance in the vertical direction. In **b**, this movement is exploited and the encoding is achieved partly by the drive field. The total measurement time was about 50 min, including a pure data acquisition time of 18 min for **a** and 1 min for **b**. Those spots with low intensity reflect imperfections of the object.

$2H_k/X_s$, where H_k is the a.c. field strength at which the material produces substantial higher harmonics, and X_s is the largest spatial derivative of a selection field component. A reasonable value for H_k may be obtained by equating the thermal energy with the Zeeman energy of the magnetic particles. Assuming $H_k = 0.5 \text{ mT } \mu_0^{-1}$, reachable with particles of 30 nm diameter, and the currently used $X_s = 3.4 \text{ T m}^{-1} \mu_0^{-1}$, the observed resolution is obtained.

Given that the particles have a reported diameter¹¹ of 4 nm it is remarkable that the achieved resolution is so high. For such particles, the Langevin theory¹² of magnetism would predict quite smooth magnetization curves with $H_k = 210 \text{ mT } \mu_0^{-1}$, leading to a resolution of the order of 10 cm. We therefore compared the observed performance to that of ideal particle ensembles acting according to the Langevin theory. Figure 4 shows the calculated normalized signal for particles in the range 10–40 nm as a function of the number of the higher harmonics. Additionally, the observed signal and the noise level are included. The experimental data fit well, assuming that particles of 30 nm diameter are responsible for the signal. The iron mass of the signal-generating particles represents only 3% of the total iron mass.

Thus far, the reconstruction was optimized to obtain the best possible resolution. It can also be adapted to compromise resolution while improving sensitivity. Analysing the signal and noise in Fig. 4, we estimate the current detection limit to be about $100 \mu\text{mol Fe l}^{-1}$ for a resolution of about 1 mm. This detection limit is already within the range of the allowed dosage for medical use¹³. However, improvements in magnetic tracers and recording electronics can be expected to lower the detection limit to $20 \text{ nmol Fe l}^{-1}$. The potential for improvement of the tracer can be seen in Fig. 4. The signal could be increased by at least two orders of magnitude (compare red line with dashed '40 nm' line) by a better initial composition and a particle separation process. Additionally, the electronics offers significant potential for improvement. With an optimized version, we expect an improvement in SNR of between one and two orders of magnitude even for a human-size system.

The detection limit extrapolated above can be verified via comparison with magnetic resonance imaging (MRI). If the signal is induced by an oscillating magnetization M , and the object dominates the noise ('patient noise limited'), the SNR is proportional to M but independent of frequency¹⁴. With the detection limit of $20 \text{ nmol Fe l}^{-1}$, the expected magnetization would still be about 5% of a

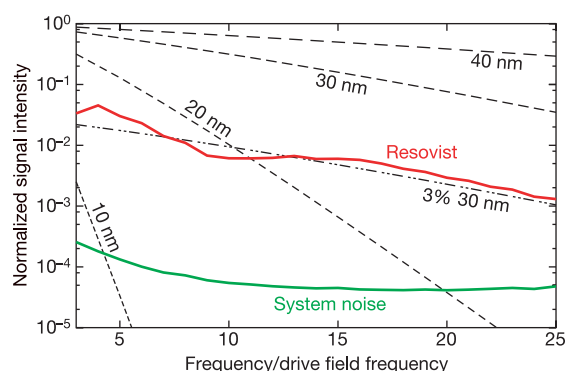


Figure 4 | Normalized signal strength as a function of frequency for simulated magnetic tracer particles and a commercially available contrast agent. The unit signal intensity is that of hypothetical maghemite particles with a step-like magnetization curve and the same iron amount as present in the commercial contrast agent, Resovist. The green line shows the measured instrumental noise of the acquisition system at a measuring time of 0.4 s. The dashed lines represent the calculated signal strength assuming a specific diameter (10–40 nm) of spherical particles acting according to the Langevin theory¹². The total amount of iron is the same as found in Resovist. Thus, the '3% 30 nm' curve represents the expected response of an assembly of ideal particles (30 nm diameter) with an iron concentration 3% that of Resovist.

typical MRI equilibrium magnetization, assuming a proton magnetization $M_{\text{MRI}}(1\text{T}) = 4 \times 10^{-9} \text{ T } \mu_0^{-1}$. So the detection limit estimated above seems to be reasonable, as the corresponding MRI signal can be detected. On the other hand, a well tolerated dosage of maghemite particles in humans¹³ is about $70 \mu\text{mol Fe l}^{-1}$, resulting in an SNR two orders of magnitude higher than that achieved using MRI.

A high SNR may speed up the image acquisition, provided that an adequate encoding speed is possible. In our experiment, the coding time was rather long, but in principle coding speed can be fast in MPI. This would require the use of three orthogonal drive fields responsible for the FFP movement (Fig. 2b), as already mentioned in the basic description. In that case, the encoding time T for a volume ($N \times N \times N$ voxel) can be shown to be roughly $T = N^2/f_1$. For a drive field frequency of $f_1 = 25 \text{ kHz}$ and $N = 50$, an encoding time of the order of 100 ms can be achieved. The size in each direction of a volume depends on the maximum possible shift $F = 2A/X_s$ of the FFP, which is of the order of 6 mm for $X_s = 3.4 \text{ T m}^{-1} \mu_0^{-1}$ and the drive field amplitude $A = 10 \text{ mT } \mu_0^{-1}$. For medical applications, a larger field of view as well as a higher coding speed is desired. Drive field amplitudes of up to $20 \text{ mT } \mu_0^{-1}$, and frequencies up to 100 kHz, may be used without harming the patient through heating. To achieve a still larger field of view, an additional, slower movement of the FFP can be superimposed. Using permanent magnets, a maximum X_s of about $3 \text{ T m}^{-1} \mu_0^{-1}$ seems to be possible for human applications, with reasonable effort. The potential of MPI becomes clear when considering the resolution and the encoding speed, in combination with the high SNR, which can be converted to imaging speed and/or sensitivity.

We have demonstrated the possibility of directly mapping magnetic material, without relying on the ill-posed inversion problem. This offers new opportunities for imaging magnetic tracers with high resolution and sensitivity. MPI may find a variety of applications, such as medical imaging, crack detection, polymer processing or fluid dynamics. For medical applications, such as vascular or small intestine imaging, the high resolution and sensitivity of MPI can be expected to be advantageous. Furthermore, the signal can penetrate tissues virtually unattenuated, allowing the inspection of regions located deep below the surface. Additionally, MPI does not necessarily require a large scanner. All required magnetic fields may be applied from one side. For a relatively small imaging volume, the scanner itself may be quite small and inexpensive. Finally, the method for localized interaction could be used not only for imaging, but also for therapy by local heating^{15,16}. Further work will be needed to exploit the full potential of this new imaging method.

METHODS

Hardware. An outline of the MPI scanner is given in Fig. 2a. The field-producing coils of the scanner are separated by 50 mm, and do not contain ferromagnetic material. In d.c. mode, the spatial derivative of the selection field X_s at the FFP is $3.4 \text{ T m}^{-1} \mu_0^{-1}$. The drive field amplitude is $10 \text{ mT } \mu_0^{-1}$. The drive field frequency is arbitrarily set to 25.25 kHz. The two recording coils are of square shape, with about 16 mm side length and a separation of about 16 mm. They are surrounded by larger windings with opposite sense (not shown in the figure) to compensate the induced voltage due to the drive field. In addition, the recorded signal is passed through a passive notch filter. After amplification and subsequent filter stages, the signal is digitized (12 bit 20 MHz, type PCI-9812, Adlink Inc.).

A robot (Flachbattanlage 1, Iselautomation KG) is used to move the sample from the upper left to the lower right corner of a square region (approximately parallel to the page plane in Fig. 2a) in subsequent horizontal lines. The resulting 52×52 data points cover a $9.4 \times 9.4 \text{ mm}^2$ region.

The first test object was the letter 'P', formed by 13 holes (diameter 0.5 mm, length 1 mm) in a flat plastic plate that were filled with the magnetic tracer. In addition, a single hole (reference object) with the same dimensions was filled with tracer and used as a reference response of the entire system. Its measurement, and later use in reconstruction, accounts for all imperfections of the coils and the complex behaviour of the tracer. Both the 'P' and the reference object were measured with identical parameters (field amplitudes, frequency, robot path, delay and recording times).

Reconstruction principles. The two acquired sets of data ('P' and reference response) were both used for reconstruction. The n th harmonic $V_n(\mathbf{y})$ of the induced signal, at the robot position vector $\mathbf{y}$, can be written as:

$$V_n(\mathbf{y}) = \int G_n(\mathbf{x} + \mathbf{y}) C(\mathbf{x}) d\mathbf{x} \quad (1)$$

Here $C(\mathbf{x})$ is the magnetic particle concentration in the object, being unknown for the image reconstruction. $G_n(\mathbf{r})$ denotes the delta response of the system, representing the induced signal in the n th harmonic of the set-up if an infinitesimally small object is placed at position $\mathbf{r}$. This function includes all the complex dynamics of the magnetic tracer, as well as the shape of the drive field and the recording coils.

Obtaining the delta response from the reference response. Owing to the use of a relatively large object (0.5 mm diameter), it is necessary to deconvolute the delta response from the reference response before starting with the reconstruction. In order to achieve that, the spatial Fourier transform of equation (1) is used (functions in Fourier space are denoted as the corresponding lower case letters with caret):

$$\hat{v}_n(\mathbf{k}) = \hat{g}_n(\mathbf{k}) \hat{c}^*(\mathbf{k}) \quad (2)$$

After division by the known concentration function $\hat{c}^*(\mathbf{k})$ of the reference object, a Fourier back-transformation yields $G_n(\mathbf{r})$.

Matrix inversion. For image reconstruction, a direct inversion of the discretized equation (1) was used, as it gives more flexibility with respect to data reduction. The induced signal is hence:

$$V_n(\mathbf{y}_i) = \delta^2 \sum_{j \in \{52 \times 52\}} G_n(\mathbf{x}_j + \mathbf{y}_i) C(\mathbf{x}_j) \quad (3)$$

Here, $\mathbf{x}_j$ and $\mathbf{y}_i$ refer to different measuring positions within the scanning plane, and δ^2 is a normalization factor due to the discretization. The concentration $C(\mathbf{x})$ in equation (1) was determined numerically for each harmonic n separately, using a zero order regularization scheme¹⁷. This leads to several complete images of the object. Finally, the mean value over a set of these individual concentration images is computed, and shown in Fig. 3a. The grey scale was assigned to the image in a linear way. Black was assigned to the lowest obtained concentration.

Reduced data set inversion. If the robot moves only in one horizontal line, the matrix equation (1) is no longer overdetermined, and all reasonable higher harmonics have to be included in the matrix to be used for inversion. The new equation can be written as:

$$\begin{aligned} V_{n(\min)} &= \delta^2 \sum_{j \in \{1 \times 52\}} G_{n(\min)}(\mathbf{x}_j + \mathbf{y}_i) C(\mathbf{x}_j) \\ &\vdots \\ V_{n(\max)} &= \delta^2 \sum_{j \in \{1 \times 52\}} G_{n(\max)}(\mathbf{x}_j + \mathbf{y}_i) C(\mathbf{x}_j) \end{aligned} \quad (4)$$

The inversion was performed in same way as above, but for each of the three lines independently. Finally, the three images were averaged to form Fig. 3b.

Received 10 January; accepted 12 May 2005.

1. Hevesy, G. Radioelements as tracers in physics and chemistry. *Chem. News J. Phys. Sci.* **108**, 166–167 (1913).
2. Haschek, E. & Lindenthal, O. A contribution to the practical use of photography according to Roentgen. *Wien. Klin. Wschr.* **9**, 63–64 (1896).
3. Fye, W. B. The coronary angiogram and its seminal contribution to cardiovascular medicine over five decades. *Circulation* **106**, 752–756 (2002).
4. Hendrick, R. E. & Haacke, E. M. Basic physics of MR contrast agents and maximization of image contrast. *J. Magn. Reson. Imaging* **3**, 137–148 (1993).
5. Goldberg, B. B., Liu, J. B. & Forsberg, F. Ultrasound contrast agents: a review. *Ultrasound Med. Biol.* **20**, 319–333 (1994).
6. Czernin, J. & Phelps, M. E. Position emission tomography scanning: current and future applications. *Annu. Rev. Med.* **53**, 89–112 (2002).
7. Weissleder, R. & Ntziachristos, V. Shedding light onto live molecular targets. *Nature Med.* **9**, 123–128 (2003).
8. Heyn, C., Bowen, C. V., Rutt, B. K. & Foster, P. J. Detection threshold of single SPIO-labelled cells with FIESTA. *Magn. Reson. Med.* **53**, 312–320 (2005).
9. Nunn, A. D., Linder, K. E. & Tweedle, M. F. Can receptors be imaged with MRI agents? *Q. J. Nucl. Med.* **41**, 155–162 (1997).
10. Romanus, E. et al. Magnetic nanoparticle relaxation measurement as a novel tool for in vivo diagnostics. *J. Magn. Magn. Matter* **252**, 387–389 (2002).
11. Lawaczek, R. et al. Magnetic iron oxide particles coated with carboxydextran for parenteral administration and liver contrasting. Pre-clinical profile of SH U555A. *Acta Radiol.* **38**, 584–597 (1997).
12. Chikazumi, S. & Charap, S. H. *Physics of Magnetism* (Wiley, New York, 1964).
13. Weissleder, R. et al. Superparamagnetic iron oxide: pharmacokinetics and toxicity. *Am. J. Roentgenol.* **152**, 167–173 (1989).
14. Vlaardingerbroek, M. T. & den Boer, J. A. *Magnetic Resonance Imaging* Ch. 6.4 (Springer, Berlin, 2003).
15. Gleich, B. Method for local heating by means of magnetic particles. World patent WO200418039 (4 March 2004).
16. Jordan, A. et al. Effects of magnetic fluid hyperthermia (MFH) on C3H mammary carcinoma *in vivo*. *Int. J. Hyperthermia* **13**, 587–605 (1997).
17. Press, W. H., Teukolsky, S. A., Vetterling, W. T. & Flannery, B. P. *Numerical Recipes in C* (Cambridge Univ. Press, 1992).

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to B.G. (bernhard.gleich@philips.com) or J.W. (juergen.weizenecker@philips.com).

LETTERS

Remobilization of southern African desert dune systems by twenty-first century global warming

David S. G. Thomas¹, Melanie Knight² & Giles F. S. Wiggs¹

Although desert dunes cover 5 per cent of the global land surface and 30 per cent of Africa, the potential impacts of twenty-first century global warming on desert dune systems are not well understood¹. The inactive Sahel and southern African dune systems, which developed in multiple arid phases since the last interglacial period², are used today by pastoral and agricultural systems^{3,4} that could be disrupted if climate change alters twenty-first century dune dynamics. Empirical data and model simulations have established that the interplay between dune surface erodibility (determined by vegetation cover and moisture availability) and atmospheric erosivity (determined by wind energy) is critical for dunefield dynamics⁵. This relationship between erodibility and erosivity is susceptible to climate-change impacts. Here we use simulations with three global climate models and a range of emission scenarios to assess the potential future activity of three Kalahari dunefields. We determine monthly values of dune activity by modifying and improving an established dune mobility index⁶ so that it can account for global climate model data outputs. We find that, regardless of the emission scenario used, significantly enhanced dune activity is simulated in the southern dunefield by 2039, and in the eastern and northern dunefields by 2069. By 2099 all dunefields are highly dynamic, from northern South Africa to Angola and Zambia. Our results suggest that dunefields are likely to be reactivated (the sand will become significantly exposed and move) as a consequence of twenty-first century climate warming.

Dune sand transport is significantly inhibited or prevented by vegetation cover of over 14% (ref. 7). Cover varies with position on the dune—flanks are normally better vegetated than crests owing to soil moisture distributions and degree of exposure to wind events. There is also significant short-term variation in subtropical dune system dynamics because of high interannual rainfall variability, droughts, localized fire and human impacts⁸. These can reduce vegetation cover on crests and flanks to less than 14%, leading to aeolian (wind-borne) sediment mobilization⁵, until vegetation recovers and crestal stabilization ensues⁹. This century, global warming is widely predicted to lead to reductions in net soil moisture and an increase in high-magnitude climatic events, including droughts in the subtropics¹⁰. Persistent changes in mean climatic parameters affecting soil moisture and windiness (which affect erodibility and erosivity) are most likely to markedly affect dunefield activity. Erodibility relates to vegetation and biocrusts and therefore to effective moisture ($P - E_p$, where P is precipitation and E_p is potential evapotranspiration), and erosivity relates to the aeolian transport capacity, expressed as the cube of mean wind speed, $\bar{U}^3$ (refs 11, 12).

In the southern Kalahari today, low wind energy limits potential sand transport to higher linear dune slopes and crests, where vegetation cover controls whether dune sand transport and erosion

takes place¹³. Eastern and northern dunefields are well vegetated and largely buffered from erosivity effects⁵. To assess twenty-first century dunefield dynamics we developed a methodology for using General Climate Model (GCM) data and an indexed measure of surface erodibility and erosivity. With a methodology established and tested, we applied it to data from a range of GCM scenarios for the three central southern African Kalahari dunefields (Fig. 1).

Monthly GCM outputs were used to assess future changes in intra-annual dune activity. Several indices of dune mobility exist that include erodibility and erosivity parameters defined by climate data^{14–16}. $M = W/(P/E_p)$, where W is the percentage of the time wind is above the threshold for sand transport, was first applied in the Kalahari⁶, with subsequent applications and calibration in other

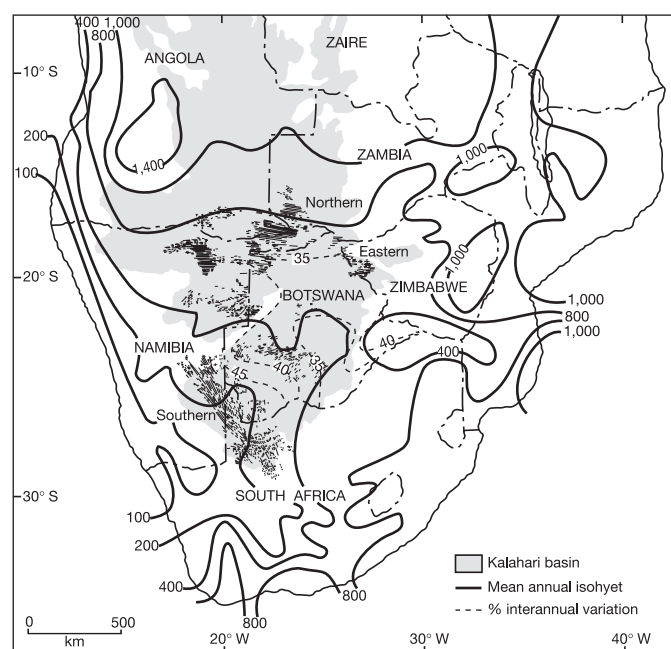


Figure 1 | The southern African Kalahari basin and dune systems. Mean annual rainfall and interannual rainfall variability isolines are shown. Most rainfall occurs October–April. In the driest, southern, areas dunes have a partial vegetation cover that responds to the high interannual rainfall variability. Low wind energy today limits sand transport and vegetation recovers after dry periods to impart stability, except where land-use pressures create bare hotspots of activity. In northern and eastern areas dunes are heavily vegetated, including supporting mixed deciduous woodland in places, owing to higher precipitation levels. Modelled 1961–90 mean A_p , GCM values correctly predict dune inactivity throughout the Kalahari. Isohyet, lines of equal precipitation.

¹School of Geography and Environment, Oxford University Centre for the Environment, University of Oxford, South Parks Road, Oxford OX1 3QY, UK. ²Department of Earth and Life Sciences, University of Salford, Manchester M5 4WT, UK.

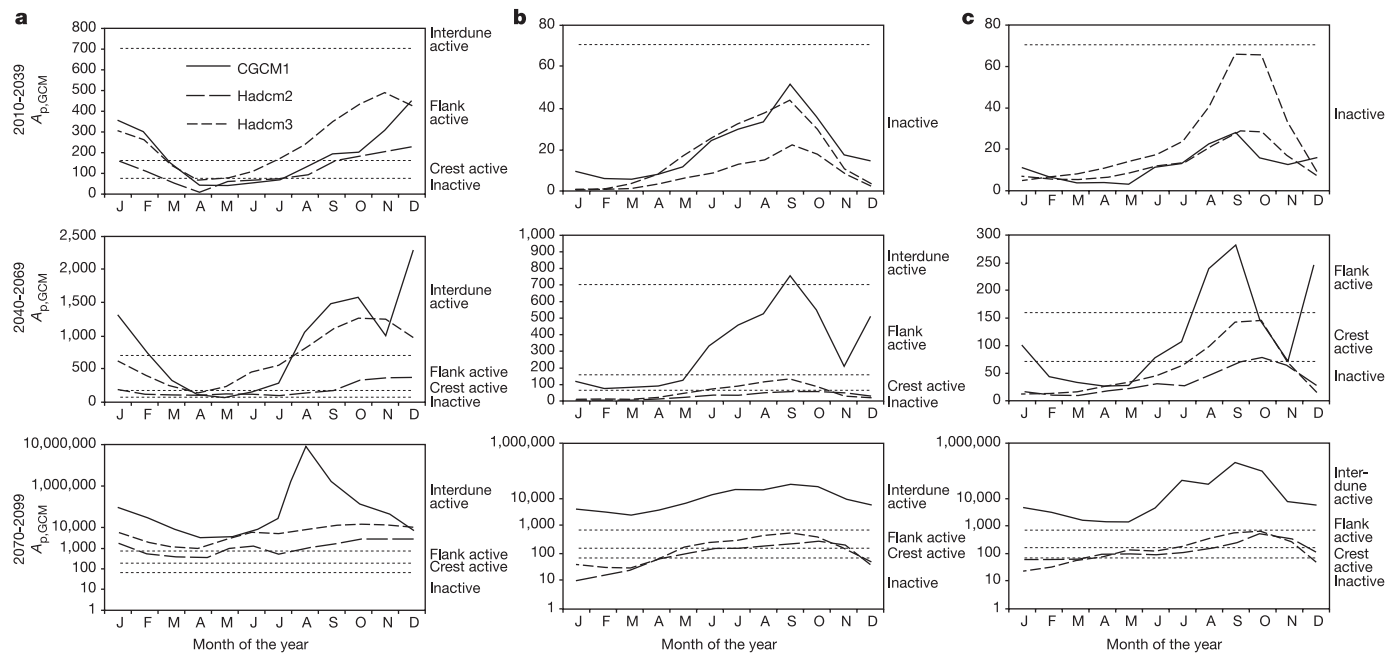


Figure 2 | $A_{p,GCM}$ dune mobility predictions using IS92a ($2 \times CO_2$) emissions scenarios. Data are shown as average annual monthly activity within each tridecadal period. **a**, Southern dunefield; **b**, northern dunefield; **c**, eastern dunefield. By 2070, dune flank activity is predicted by all model outcomes in the southern dunefield (bottom panel of **a**). Significantly,

interdune activity, indicating a fully active dunefield, is predicted in two scenarios. In the northern Kalahari, significant activity is also predicted after 2040 (middle panel of **b**), suggesting a marked change in landscape stability and ecosystems.

dunefields^{17,18}. We modified the index M to account for GCM data outputs and to determine monthly values, improving sensitivity in a region of highly seasonal climate (see Methods). GCMs were evaluated to provide twenty-first century climate predictions, with consideration given to grid sizes, ability to postdict the 1961–90 climate for index validation, and sensitivity to atmospheric composition and level of greenhouse gases¹⁹. Hadcm3, Hadcm2 and CGCM1 were selected for their relatively fine resolutions, with predictions of future global warming ranging from 2.5°C (Hadcm3) to 3.5°C (CGCM1) by 2100. These GCMs allowed different model generations to be incorporated into the study. Testing indicated that no one GCM performed well for all key variables, but for Hadcm3 and Hadcm2, values of the revised index of potential dune activity, $A_{p,GCM}$ (see Methods for definition) bore a close relationship to those derived from observed data for the 1961–90 period. Monthly and annual GCM data runs were used for future climate scenarios, with results averaged over tridecadal blocks (2010–2039, 2040–2069 and 2070–2099). We used the IS92a ($2 \times CO_2$) emission scenario and several SRES scenarios to account for increases in additional atmospheric gases²⁰. Once predictions were made for the erosivity and erodibility elements of $A_{p,GCM}$, the

results were analysed to consider the impacts of spatial and seasonal²¹ trends in future climate.

Using IS92a, all the GCMs produced twenty-first century climate scenarios leading to increased dunefield erodibility, though with differing changes in P and E_p . Csiro-mk2b predicts a doubling of E_p throughout southern Africa by 2100. The older CGCM1 predicts a 50% decline in summer rainfall in northern areas, accompanied by a quadrupling of E_p . Although the newer Hadley GCMs predict smaller increases in E_p than the other models, the P/E_p ratio was always <1 . This negative moisture budget applies even to the currently heavily vegetated northern dunefield, for which using SRES emission scenarios Hadcm3 predicts a 50% precipitation increase. When compared, the IS92a and SRES results agree, in that any modelled annual precipitation increases in the southern African interior are outweighed by evapotranspiration increases. Some changes in rainfall seasonality are modelled, but only Csiro-mk2b with SRES scenarios suggests a heightening of dry–wet season contrasts.

Erosivity is projected to increase in all model and scenario outputs, with greatest increases in the southern dunefield. The most sensitive outputs are generated by Hadcm3 with both IS92a and SRES. Ramped erodibility increases are driven by a doubling of the present

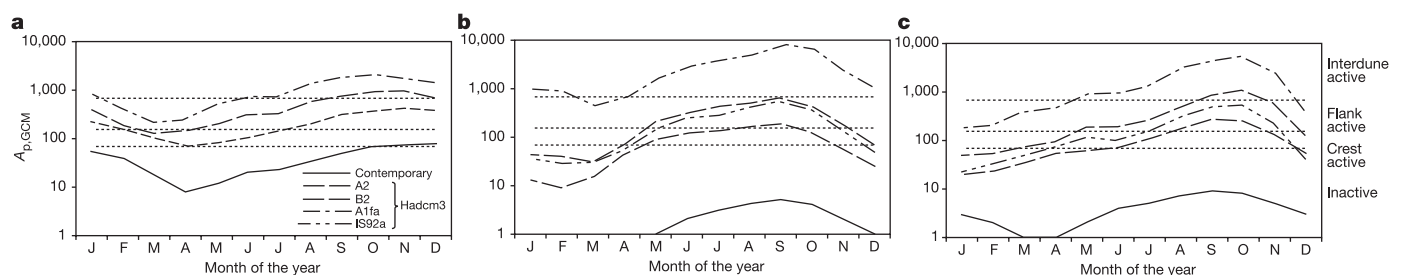


Figure 3 | Contemporary and post 2070 $A_{p,GCM}$ values of Kalahari dunefield activity. **a**, Southern dunefield; **b**, northern dunefield; **c**, eastern dunefield. ‘Contemporary’ indicates values calculated using 1961–90 climate data. Post-2070 runs with the Hadcm3 GCM use a range of emission

scenarios: A2, SRES ‘business as usual’ emissions; B2, SRES medium emissions scenario; A1fa, SRES emissions with increasing fossil fuel use; IS92a, the $2 \times CO_2$ scenario.

mean wind speeds to 4 m s^{-1} after 2040. Little change occurs in the timing of wind-speed maxima throughout the year, with the focus remaining in the August–October period.

To establish potential future dunefield activity we integrated temporal determinations of erodibility and erosivity from specific GCM and emission scenarios into $A_{p,GCM}$. Under the present climate conditions, only the southern dunefield experiences activity and this is largely limited to dune crests during droughts⁷. All modelled outputs project marked increases in dune activity during the twenty-first century in all dunefields, including after 2040 in the northern dunefield (Botswana, Namibia, Angola) and in the eastern dunefield (Zimbabwe, Zambia). Significantly, by 2070 activity levels even in northern areas exceed the current activity maxima in the dry southwestern areas. No particular GCMs or emission scenarios consistently produce the 'least active' or 'most active' predictions, but with IS92a emissions CGCM1 often produces the most active scenarios (Fig. 2).

The GCMs all produce increasing levels of activity over the twenty-first century, including fully active dunefields, where interdune areas are devegetated and sand mobility and transfer between dunes is possible, resulting in marked landscape changes such as plant community diminutions. These would represent considerable, even catastrophic, limitations on the present agricultural uses of these environments. Our model does not account for the potential impact of enhanced atmospheric CO_2 levels on plant productivity, which may be greatest in dry environments²². These might counter some

of the cover reductions due to reduced future net precipitation; however, critical advantages may be lost under the impacts of droughts that are expected to increase in severity and frequency²³. From the middle of the twenty-first century projected values of $A_{p,GCM}$ are exceedingly high ($>5,000$ in many months with several GCM runs) owing to both substantial moisture depletion and markedly enhanced windiness.

We evaluated differing emission scenario impacts on potential dune activity (Fig. 3). Although the extreme situation generated by the use of the SRES A1fa scenario is notable, perhaps of greater significance is that even the medium-emission scenario, B2, generates significant aeolian dynamism across the Kalahari, including the currently heavily vegetated eastern and northern dunefields. The limited southern dunefield activity today displays year-to-year fluctuations in association with climate variability⁹; interannual variability in dune activity levels is modelled to continue in the late twenty-first century too, but even in the least active years $A_{p,GCM}$ values are sufficiently high that dune crests would remain active. If these predictions were to be correct, there would be little respite from sand mobility in the Kalahari dunefields even during relatively moist years.

Finally, we took $A_{p,GCM}$ outputs for the post-2070 period using Hadcm3 and determined the mean status of runs using all four of the emission scenarios used in this analysis (Fig. 4). This indicates that the Kalahari has the potential to achieve levels of aeolian activity that have not operated since 14–16 kyr ago—the last period of

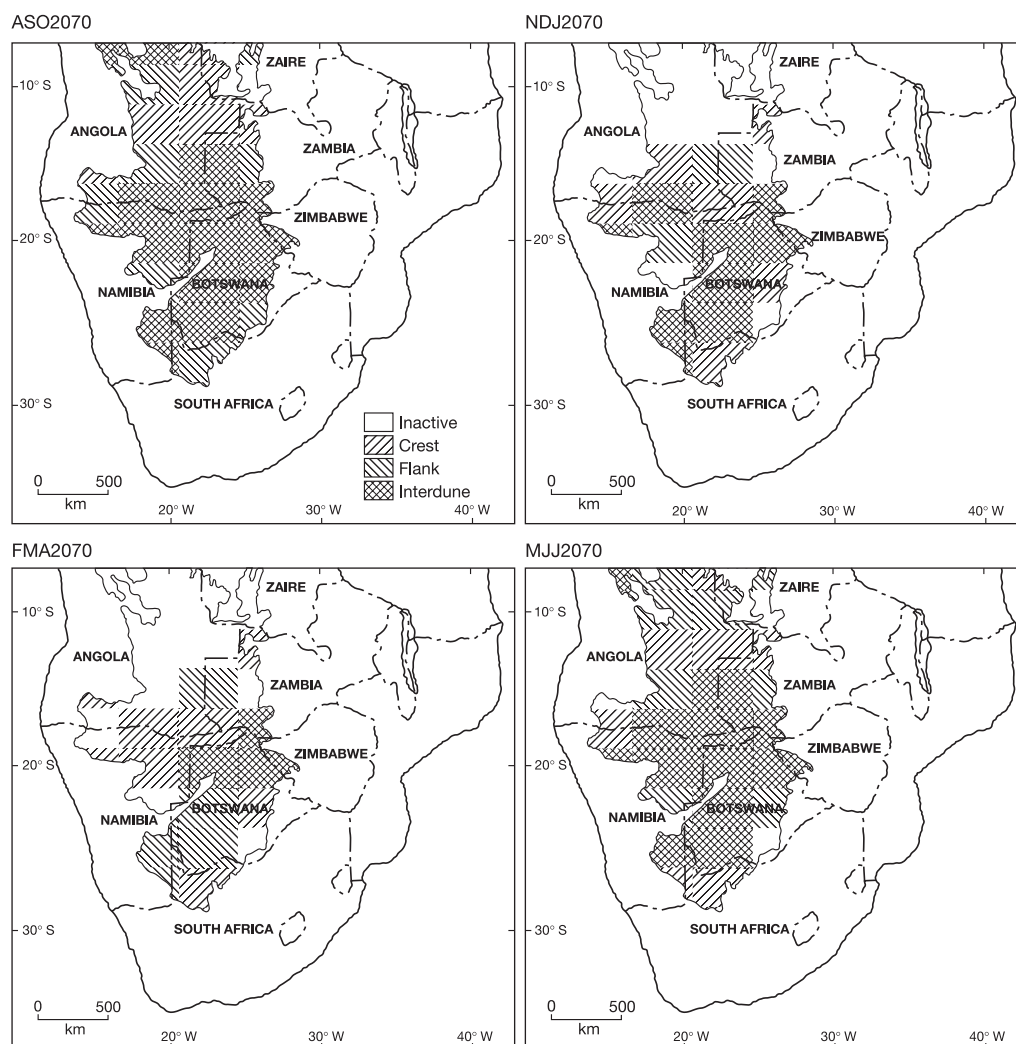


Figure 4 | Predicted three-month block dunefield activity after 2070. This is modelled as the mean of Hadcm3 runs using IS92a, A2, A1fa and B2

emission scenarios. Significant dunefield activity occurs all year round in all three dunefields.

Kalahari-wide dunefield activity recorded in the luminescence-dated sedimentary record²⁴.

Our findings show that dunefields are likely to experience significant reactivations as a consequence of twenty-first century climate change. This finding is independent of the GCM or emission scenario used. There are uncertainties within the modelled Kalahari scenarios but the general trend and the magnitude of possible changes in the erodibility and erosivity of dune systems suggests that the environmental and social consequences of these changes will be drastic.

METHODS

Potential activity index. To adapt M^6 to GCM data and apply it to a highly seasonal climate regime, several changes were made. Because mean monthly wind velocity, U (in m s^{-1}), is the common GCM wind output, W was replaced by $\bar{U}^3$, which has been employed in other dune mobility indices^{14,15}, and is widely used in agricultural wind erosion models^{11,25}, and in investigations of short-event-based sand flux¹². At present vegetated dunes extend 10° latitude north of the arid southern Kalahari, where M was developed; these northerly locations today receive up to 900 mm of precipitation per year. The moisture retention capacity of dune bodies²⁶ allows for the more persistent impact of prolonged wet (and dry) periods on soil moisture and vegetation cover through the year²⁷. A first-approximation weighting for rainy season net precipitation was therefore introduced, as well as one for the preceding two months rainfall¹⁶. This weighting is also likely to have a conservative effect on predicted mobility, through its effect on erodibility. The revised index is $A_{p,GCM} = \bar{U}^3 / (P_{lag}/E_{p,lag} + P_{rainy}/E_{p,rainy})$, where $\bar{U}^3$ = the cube of the mean wind speed. $P_{lag}/E_{p,lag}$ is the residual effect of recent rainfall and potential evaporation, such that $P_{lag} = (P_{-1} + P_0)/2$, where P_{-1} is precipitation in the previous month and P_0 is rainfall in the current month, and $E_{p,lag} = (E_{p,-1} + E_{p,0})/2$, where $E_{p,-1}$ is potential evapotranspiration in the previous month and $E_{p,0}$ is potential evapotranspiration in the current month. $P_{rainy}/E_{p,rainy}$ is the effect of rainy season precipitation and potential evaporation on soil moisture, such that $P_{rainy} = (P_N + P_D + P_J...)/m$ and $E_{p,rainy} = (E_{p,N} + E_{p,D} + E_{p,J}...)/m$, where $m = N$ (November), D (December), J (January), and so on (abbreviations as in the figures) is the month under consideration within the rainy season. $A_{p,GCM}$ is adapted for the limitations of GCM data outputs, is specifically developed for the seasonal nature of southern African climates, and contains determinations of erodibility and erosivity elements that are not expected to overestimate potential dune dynamics.

Model validation. For dune activity indices the thresholds between classes are important for model calibration. A fourfold division of $A_{p,GCM}$ values was achieved through a validation exercise that explains different levels of aeolian activity within dunefields: $A_{p,GCM} > 700$ indicates highly dynamic dune landscapes, bare dune bodies and sparsely vegetated interdunes; $A_{p,GCM} = 160\text{--}700$ indicates significant dune activity, dunes crests bare, dune flanks rippled but moderate interdune vegetation cover; $A_{p,GCM} = 70\text{--}160$ indicates dune activity limited to crests, dune flanks vegetated, interdunes well vegetated; and $A_{p,GCM} < 70$ indicates vegetation cover across the whole dune, dunes inactive. Contemporary tridecadal (1961–90) $A_{p,GCM}$ values for the southern dunefield correctly indicate mean dunefield inactivity (Fig. 3); for the northern dunefield $A_{p,GCM} < 10$ for all months, < 50 for the eastern dunefield and up to 100 for the southern dunefield. Within these data, high interannual climatic variability (Fig. 1) and rainfall seasonality leads to variations in Kalahari activity status^{9,16}.

Validation was achieved using monthly $A_{p,GCM}$ values from 1960–2000 climatic data from Twee Riverien in the southern dunefield, calibrated against temporally referenced empirical dune activity and vegetation cover data from published and field research sources. $A_{p,GCM}$ values from $> 9,000$ to 1 were calibrated against: (1) empirical vegetation cover data for September 1999 and June 2000; (2) dune surface activity data based on eight years (1992–2000) repeat dry-season measurements and observations at a number of sites within 10 km of the meteorological station, including published data for 1992 (ref. 7); (3) a photographic library of dune images from 1983–2000 providing further evidence of vegetation conditions on a range of linear dune and interdune areas; and (4) remote sensing and field-based analyses of vegetation cover in the dunefields^{28,29}, including detailed analyses of two Thematic Mapper satellite image subscenes centred on Twee Riverien in the southern dunefield (1984 and 1993)⁹. Supplementary Fig. 1 illustrates and exemplifies the calibration. In addition, monthly $A_{p,GCM}$ values from 2001–2 automatic weather station data in the very arid, active, negligibly vegetated Namib dunefield, always exceed 31×10^6 . Projected values for the southern Kalahari dunefield attain 88×10^6 for August in 2070–2099 with one GCM scenario (Fig. 2).

Received 7 October 2004; accepted 3 May 2005.

1. Thomas, D. S. G. in *Arid Zone Geomorphology* (ed. Thomas, D. S. G.) 373–412 (Wiley, Chichester, 1998).
2. Stokes, S., Thomas, D. S. G. & Washington, R. Multiple episodes of aridity in southern Africa since the last interglacial period. *Nature* **388**, 154–158 (1997).
3. Sporton, D. & Thomas, D. S. G. *Sustainable Livelihoods in Kalahari Environments. Contributions to Global Debates* (OUP, Oxford, 2002).
4. Turner, M. D. Spatial and temporal scaling of grazing impact on the species composition and productivity of Sahelian annual grasslands. *J. Arid Environ.* **41**, 277–297 (1999).
5. Livingstone, I. & Thomas, D. S. G. in *The Dynamics and Context of Aeolian Systems* (ed. Pye, K.) 91–101 (Geol. Soc. Spec. Pub. 72, Geological Society, London, 1993).
6. Lancaster, N. Development of linear dunes in the southwestern Kalahari, southern Africa. *J. Arid Environ.* **14**, 233–244 (1988).
7. Wiggs, G. F. S., Thomas, D. S. G., Bullard, J. E. & Livingstone, I. Dune mobility and vegetation cover in the southwest Kalahari Desert. *Earth Surf. Proc. Landforms* **20**, 515–529 (1995).
8. Wasson, R. J. & Nanninga, P. M. Estimating wind transport of sand on vegetated surface. *Earth Surf. Proc. Landforms* **11**, 505–514 (1986).
9. Thomas, D. S. G. & Leason, H. Dunefield activity response to climate variability in the southwest Kalahari. *Geomorphology* **64**, 117–132 (2005).
10. International Panel on Climate Change (IPCC). *Climate Change 2001* (Cambridge Univ. Press, Cambridge, 2001).
11. Woodruff, W. P. & Armbrust, D. V. A monthly climatic factor for the wind erosion equation. *J. Soil Wat. Conserv.* **23**, 103–104 (1968).
12. Bauer, B. O., Yi, J., Namikas, S. L. & Sherman, D. J. Event detection and conditional averaging in unsteady aeolian systems. *J. Arid Environ.* **39**, 345–375 (1998).
13. Wiggs, G. F. S., Thomas, D. S. G., Bullard, J. E. & Livingstone, I. Dune mobility and vegetation cover in the southwest Kalahari Desert. *Earth Surf. Proc. Landforms* **20**, 515–529 (1995).
14. Talbot, M. R. Late Pleistocene rainfall and dune building in the Sahel. *Palaeoecol. Afr.* **16**, 203–214 (1984).
15. Kar, A. Aeolian processes and bedforms in the Thar Desert. *J. Arid Environ.* **25**, 83–96 (1993).
16. Bullard, J. E., Thomas, D. S. G., Livingstone, I. & Wiggs, G. F. S. Dunefield activity and interactions with climatic variability in the southwest Kalahari desert. *Earth Surf. Proc. Landforms* **22**, 165–174 (1997).
17. Muhs, D. R. & Maat, P. B. The potential response of eolian sands to greenhouse warming and precipitation reduction on the Great Plains of the USA. *J. Arid Environ.* **25**, 905–918 (1993).
18. Lancaster, N. & Helm, K. A test of a climatic index of dune mobility using measurements from the southwestern United States. *Earth Surf. Proc. Landforms* **25**, 197–207 (2000).
19. Knight, M., Thomas, D. S. G. & Wiggs, G. F. S. Challenges of calculating dunefield mobility over the 21st century. *Geomorphology* **59**, 197–213 (2004).
20. Nakicenovic, N. & Swart, R. *IPCC Special Report on Emissions Scenarios* (IPCC, UNEP & WMO, Cambridge Univ. Press, Cambridge, 2000).
21. Summer, G., Homar, V. & Ramis, C. Precipitation seasonality in eastern and southern coastal Spain. *Int. J. Clim.* **21**, 211–247 (2001).
22. Körner, C. Biosphere responses to CO₂ enrichment. *Ecol. Appl.* **10**, 1590–1619 (2000).
23. Hulme, M. *Climate Change and Southern Africa* (CRU, Univ. of East Anglia, Norwich, 1996).
24. Thomas, D. S. G. & Shaw, P. A. Late Quaternary environmental change in central southern Africa: new data, synthesis, issues and prospects. *Quat. Sci. Rev.* **21**, 783–798 (2002).
25. Chepil, W. S. & Woodruff, N. P. The physics of wind erosion and its control. *Adv. Agron.* **15**, 211–302 (1963).
26. Tsoar, H. & Möller, J. T. in *Aeolian Geomorphology* (ed. Nickling, W. G.) 75–95 (Allen and Unwin, Boston, 1986).
27. Scanlon, T. M., Albertson, J. D., Caylor, K. K. & Williams, C. A. Determining land surface fractional cover from NDVI and rainfall time series for a savanna ecosystem. *Remote Sens. Environ.* **13**, 419–428 (2002).
28. Ringrose, S. & Mathieson, W. A Landsat analysis of range conditions in the Botswana Kalahari drought. *Int. J. Remote Sens.* **12**, 1023–1051 (1991).
29. Nicholson, S. E. & Kim, J. The relationship of the El Niño oscillation to African rainfall. *Int. J. Clim.* **17**, 117–135 (1997).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank P. Coles for draughting figures, and the University of Sheffield for financial support to M.K. H. Viles supplied Namib weather data used in model validation.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to D.S.G.T. (david.thomas@ouce.ox.ac.uk).

LETTERS

Imaging the Indian subcontinent beneath the Himalaya

Vera Schulte-Pelkum¹, Gaspar Monsalve¹, Anne Sheehan¹, M. R. Pandey², Som Sapkota², Roger Bilham¹ & Francis Wu³

The rocks of the Indian subcontinent are last seen south of the Ganges before they plunge beneath the Himalaya and the Tibetan plateau. They are next glimpsed in seismic reflection profiles deep beneath southern Tibet¹, yet the surface seen there has been modified by processes within the Himalaya that have consumed parts of the upper Indian crust and converted them into Himalayan rocks^{2,3}. The geometry of the partly dismantled Indian plate as it passes through the Himalayan process zone has hitherto eluded imaging. Here we report seismic images both of the decollement at the base of the Himalaya and of the Moho (the boundary between crust and mantle) at the base of the Indian crust. A significant finding is that strong seismic anisotropy develops above the decollement in response to shear processes that are taken up as slip in great earthquakes at shallower depths. North of the Himalaya, the lower Indian crust is characterized by a high-velocity region consistent with the formation of eclogite, a high-density material whose presence affects the dynamics of the Tibetan plateau.

In 2001–03, we operated 29 broadband seismometers in Nepal and Tibet for 18 months (Fig. 1), during which time we located ~1,700 earthquakes within the region. All distant earthquakes recorded were subjected to an automated selection process for receiver function calculation (by magnitude, distance, signal-to-noise ratio of the P arrival, and variance reduction of the receiver function). Removal of a handful of remaining outliers during visual inspection left ~40–250 high-quality receiver functions per station. Receiver function analysis allows determination of the time delay between near-vertically travelling direct P waves, and converted S waves that travel to the seismometer from subsurface interfaces with velocity contrast. Delays between the direct and converted wave are proportional to the depth of the interface and depend on the transmission velocities along their paths, while the amplitude of the converted arrival depends on the magnitude and sign of the velocity contrast. The delay times can be converted to interface depths assuming a velocity model. In Fig. 2b, we present a subsurface profile produced by migration and geographical stacking of the receiver functions using methods^{4,5} similar to those developed for reflection seismology. Tests show that the structure is sufficiently uniform along the Himalayan arc for a single arc-normal projection within the area studied to be valid (see Supplementary Information).

The clearest feature on the processed image is the Moho, the surface separating the high velocities of the mantle from the slower ones in the Indian continental crust. The Moho appears as a near-horizontal surface beneath India (at ~45 km depth) and Tibet (at ~75 km depth), and is offset smoothly downward beneath the Himalaya over a distance of 120 km. The base of the Indian crust is thus well determined by the receiver function imagery, and confirms

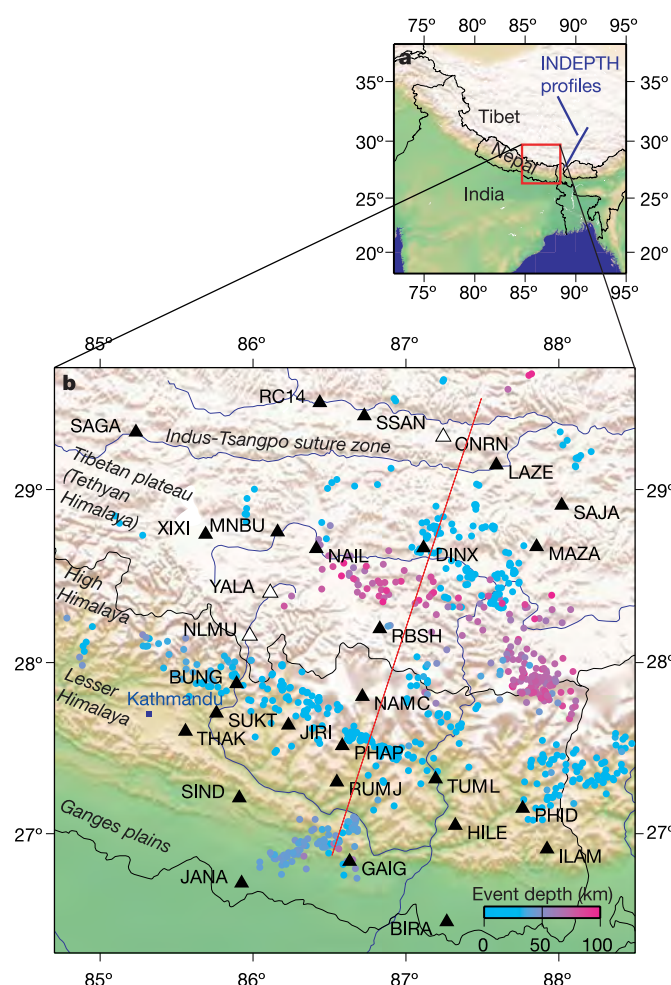


Figure 1 | Location map. **a**, Overview map with topography. The extent of the study area map in **b** is outlined in red. The location of INDEPTH profiles¹ is indicated in blue. **b**, Topography map of the study area. Stations deployed for this study are shown in black (three stations with little to no data owing to equipment problems or vandalism are shown in white). Hypocenters relocated with our network are colour coded by depth (scale in km). The red line is the location of the profile in Fig. 2 with end points 26.873° N, 86.517° E, and 29.525° N, 87.495° E, orientation N18E. Southern stations BIRA, JANA and GAIG are situated on thick sediments whose multiples dominate their receiver functions and are therefore not used for stacks shown in Fig. 2.

¹Cooperative Institute for Research in Environmental Sciences (CIRES), University of Colorado, Boulder, Colorado 80309-0399, USA. ²Department of Mines and Geology, Lainchaur, Kathmandu, Nepal. ³SUNY, Binghamton, New York 13902, USA.

widely-held views that the lower Indian crust underplates the southernmost Tibetan plateau.

Stations south of the High Himalaya delineate a shallow layer, whose seismic velocity is strongly dependent on incoming wave direction (Fig. 3a, c; see also Supplementary Information). Azimuthal variations in receiver functions can be caused by lateral variations, but several features of these early arrivals—the polarity reversal on the radial component, the strong transverse component energy without a preceding arrival at zero time, and repetition of the pattern at stations distributed over an extended area—require the presence of an anisotropic layer. Waves emerging steeply from the north experience an increase in velocity entering this layer, creating a negative polarity conversion, whereas waves emerging from the south experience a velocity decrease and show a positive polarity arrival. Their sum renders the layer invisible in the standard stack in Fig. 2b. In order to allow detection of interfaces with azimuthal polarity reversals, we calculated an azimuthal contrast stack of the difference between southern and northern azimuth arrivals (Fig. 2c), which shows clearly the base of the anisotropic layer in Nepal and a hint of its continuation under Tibet.

We can reproduce the azimuthal pattern in the receiver functions (Fig. 3b) with a strongly ($\sim 20\%$) anisotropic layer with foliation planes of fast seismic velocity dipping steeply down ($\sim 50^\circ$) towards a strike of north-northeast, the direction of plate convergence. The layer has a thickness of ~ 6 km, and its base dips from 8 km depth at the southern margin of the Nepal foothills to 20 km depth just south of the High Himalaya, coinciding with the location of the decollement separating the Himalaya and the Indian plate inferred from structural geology^{6–8}. Limitations of depth resolution and unknown subsurface velocity variations place an uncertainty of 2 km on these depths. The anisotropic fabric in the hanging wall of the Himalaya is caused by shear on the decollement, which induces development of foliation planes that concentrate minerals such as micas and amphiboles (Fig. 3d). Shear-induced foliation planes tilt steeply downward in the direction of shear⁹, consistent with our observations. Mineral alignment occurs under ductile conditions above temperatures of $\sim 250^\circ\text{C}$, conditions that can only prevail in the deeper, northern portion of the imaged shear zone, below the transition from locked to stable sliding that has been inferred on the decollement from geodetic evidence¹⁰. A concentration of small earthquakes at the depth of the

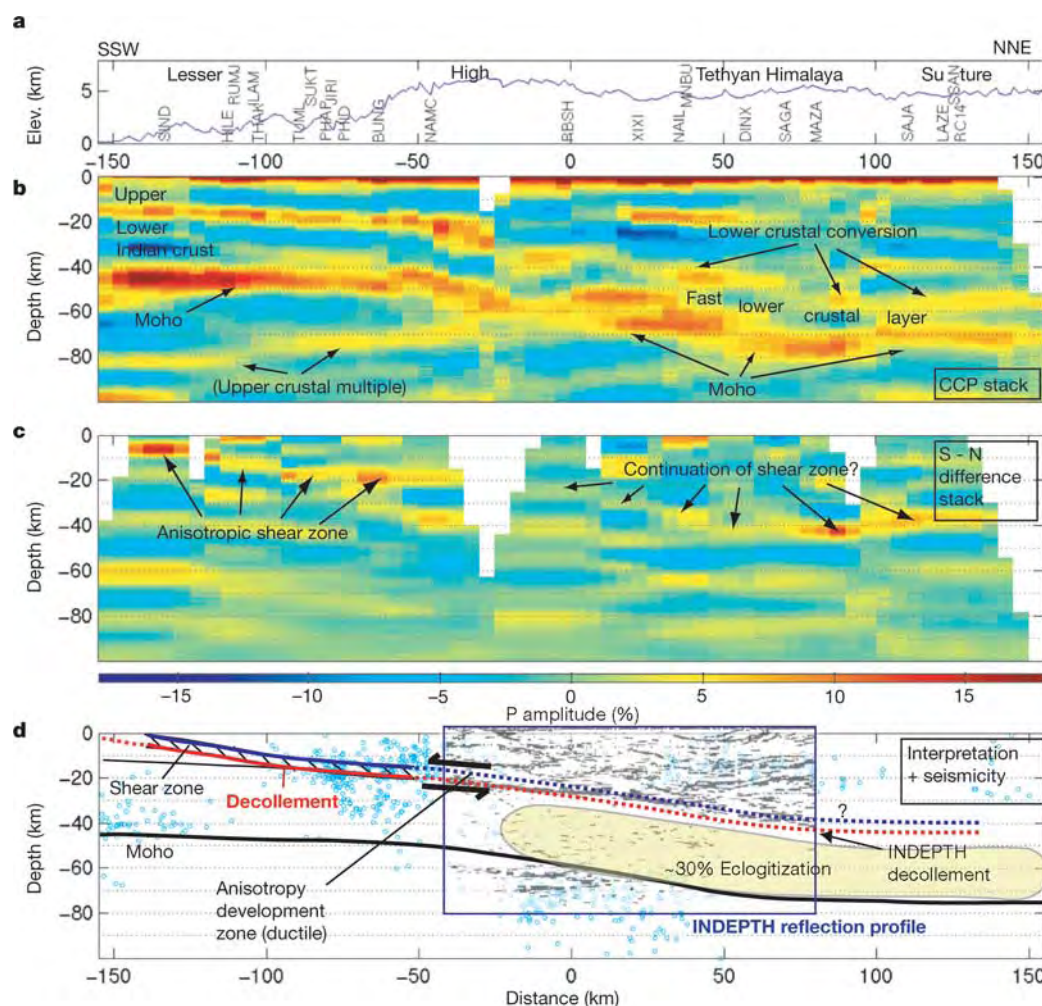


Figure 2 | Receiver function analysis along profile and interpretation.

a, Topography and station locations along the N18E profile indicated in Fig. 1. Elev., elevation. **b**, Common conversion point (CCP) stack of radial receiver functions along the same profile, migrated from time to depth with a simple three-dimensional velocity model from tomographic inversion (one one-dimensional model for Nepal, with station corrections for SIND and BUNG, and one one-dimensional model for Tibet; Supplementary Table 2). White gaps are areas with no data coverage. Bin width is 5 km, with lateral smoothing over 30 km. Colour scale for **b** and **c** is stacked receiver function amplitude expressed as a percentage of the direct P arrival. (See

Supplementary Information for discussion of the upper crustal multiple.) **c**, South-north azimuthal difference common conversion point stack: the amplitudes of receiver functions from southern versus northern azimuths are differenced before stacking to allow imaging of arrivals that change polarity over back-azimuth (see also Supplementary Information). Lateral smoothing is now limited to 10 km to allow more precise differencing of receiver functions from opposite azimuths. Scale as in **b**. **d**, Interpretation of **b** and **c**, with INDEPTH reflection profile^{1,8} and seismicity located with our network superimposed. Scale as in **b** and **c**.

decollement occurs just south and upslope of the presumed brittle-to-ductile transition, and north and downslope of the locked portion of the decollement (Fig. 2d).

The anisotropy that we observe at depths above the brittle–ductile transition is presumably conveyed upwards along the decollement

from deeper levels by tectonism and exhumation, as demonstrated by surface exposures of the Main Central thrust that have been mapped locally and shown to possess similar several-kilometres-thick shear zones¹¹. Strong anisotropy from ductile deformation indicates that significant local strain and pure shear exist in the lower part of the Himalayan wedge, which may encourage structural geologists to further investigate the role of finite strain in balanced cross-sections.

On the basis of our observed geometry and receiver function studies from India¹², we assume a midcrustal interface imaged at 17 to 20 km depth in the southern third of the profile to be an inherited feature of the Indian crust, unrelated to the collision to the north. In Fig. 2, the Moho and midcrustal interface in the Nepal foothills in the southern part of the profile are parallel to each other, with a gentle northward dip. In contrast, the decollement, indicated by the base of the anisotropic shear zone, slices down northward at a steeper angle from near the top of the upper crust down into the midcrustal interface. This geometry suggests that the upper part of the Indian crust detaches along the base of the shear zone from the deeper portion and is incorporated into the Himalaya, while the lower crust continues its descent under Tibet. The position of the decollement as we observe it under Nepal, and the hint of its possible continuation that we see under Tibet (Fig. 2c), correspond well with the decollement imaged by INDEPTH reflection data ~200–300 km east of our study area (Fig. 2d). The accumulation of only upper-crustal Indian rocks into the Himalaya is consistent with geological findings^{2,13}. The descending lower part of the crust, under increasing pressures and temperatures, appears to undergo changes in material properties.

A crustal positive amplitude arrival, seen at all Tibetan stations at depths of 45–55 km, implies a high-velocity layer in the lower crust. At the same stations, conversions from the Moho are weaker than we observe in Nepal, which is expected for a reduced velocity contrast at the Moho due to a fast lower crust. We estimate crustal velocities using arrival times of local and regional earthquakes recorded at our network¹⁴. Minimizing travel time residuals in earthquake locations, as well as tomographic inversions, require a fast lower crust under the Tibetan plateau, with P velocities of over $\sim 7.0 \text{ km s}^{-1}$ (Supplementary Figs 3, 4). In contrast, the same procedures applied to events in Nepal return a crustal velocity model with normal wave speeds very similar to published values¹⁵, with a P velocity in the lower crust of $6.4\text{--}6.5 \text{ km s}^{-1}$. Some increase in velocity would be expected as the Indian lower crust is subjected to higher pressures and initially cold temperatures at increasing depths. However, an increase in P velocity as large as $\sim 0.5 \text{ km s}^{-1}$ in lower-crustal materials as a result of a 30 km burial is unlikely without the additional influence of phase transitions, on the basis of laboratory measurements¹⁶.

A much more likely explanation for the anomalous velocity increase is partial eclogitization of the lower Indian crust under Tibet. Eclogite is seismically fast, and our observed velocities suggest that $\sim 30\%$ of the lower Indian crust undergoes this phase transition. As the density of the converted material increases by up to $\sim 21\%$ (ref. 17), the total lower-crustal volume would be reduced, and its density increased, by up to 6%. Although large-scale, pervasive

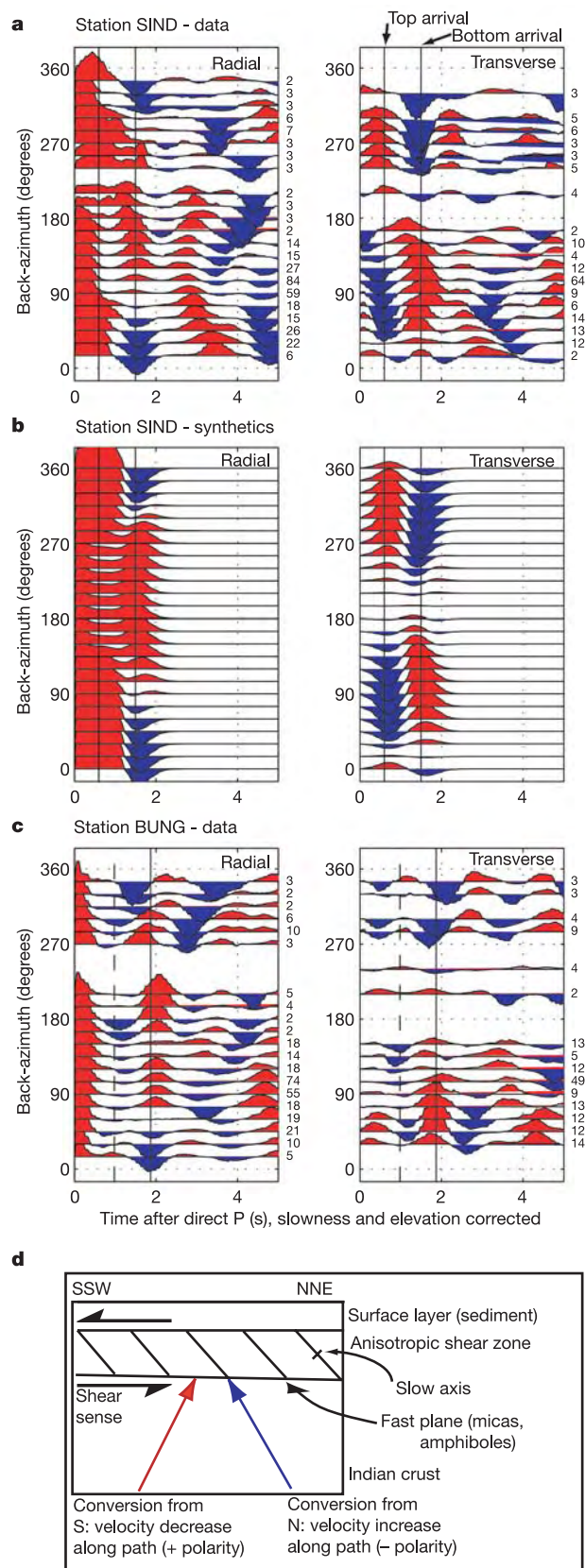


Figure 3 | Observed and synthetic data showing the anisotropic shear zone, and explanation. **a**, Radial (left) and transverse (right) receiver functions observed at southern Nepal station SIND, with number of receiver functions per 15° bin average indicated. Slight azimuthal smoothing is applied (5° overlap between bins, no duplication in count on right). Positive arrivals are shown in red, negative in blue. Arrivals with polarity reversal are indicated. **b**, Synthetic receiver functions³² (see Supplementary Information) calculated for a shallow, strongly (20%) anisotropic layer to match the polarity reversal arrivals observed at SIND. **c**, The northern Nepal station BUNG shows similar arrivals later than at SIND, indicating that the depth of the layer increases to the north. Although SIND and BUNG are the best examples, similar arrivals are seen at all Nepal stations. **d**, Model used for synthetics at SIND with explanation of the anisotropy development and the mechanism of polarity reversal in radial receiver function amplitudes.

conversion to eclogite can result in delamination of the denser material into the mantle, we see no evidence of such a process in the seismic image. The timing and amplitude of both Moho and lower-crustal interface in Tibet vary from station to station and with back-azimuth at the same station, making for a diffuse appearance in the stack compared to the Moho under Nepal (Fig. 2). This appearance suggests that eclogitization is incomplete and distributed, presumably influenced by water availability¹⁸. The lower Indian crust appears to underplate Tibet at least as far north as the Indus-Tsangpo suture. Further support for the presence of a layer containing eclogite is provided by INDEPTH results to the north and east of our study area, with findings of a Moho 'doublet' (corresponding to our lower-crustal high-velocity layer) just north of the Indus-Tsangpo suture that vanishes further north⁵.

Our seismic profile allows us to answer several outstanding questions in Himalayan mountain building. Gravity measurements predicted an increase in Moho depth under the High Himalaya^{19,20}, but whether this occurred along a steepened continuous Moho, as we confirmed in this study, or on a stepped Moho in an imbricated crust, was a matter of debate^{21–23}. We provide seismic evidence for a single decollement south of the High Himalaya, and show that strong deformation of the upper Indian crust, and its incorporation into the Himalaya, may quantitatively reconcile inequities in previous estimates of crustal volume budgets within the collision zone².

Eclogitization and loss of the denser material into the mantle have been invoked as possible mechanisms to balance the crustal volume deficit^{24,25}. However, several seismic studies have inferred abnormally low, rather than increased, velocities for the Tibetan lower crust^{24,26,27}, suggesting an absence of eclogite. Our Moho depths and the crustal velocities used to calculate depths are internally consistent owing to our joint use of receiver functions, hypocentre determination, and tomographic inversion, and we are able to confirm the presence of fast material above the Tibetan Moho, with velocities and geographical distribution that suggest partial and diffuse eclogitization. Although eclogitization has recently been proposed as an explanation²⁸ for deep earthquakes under Tibet, our internally consistent event locations and Moho depths confirm previously held views²⁹ that the deep events occur in the mantle. Eclogitization can therefore only be an indirect cause of the deep seismicity. Our observation of a fast lower crust under Tibet is the first seismic confirmation of a previously postulated zone of eclogite between the High Himalaya and the Indus-Tsangpo suture, where cold temperatures due to fast underthrusting initially inhibit the eclogite transition and allow the High Himalaya to reach its unusual elevation before eclogite forms north of the High Himalaya³⁰. As the lower crust subsequently heats up to reach temperature equilibrium, the eclogite probably converts to low-density granulite north of the Indus-Tsangpo suture (suggested also by the disappearance of the INDEPTH Moho 'doublet'⁵), which would help buoy up the Tibetan plateau³¹.

Received 9 November 2004; accepted 29 March 2005.

1. Zhao, W. & Nelson, K. D. Project INDEPTH Team. Deep seismic reflection evidence for continental underthrusting beneath southern Tibet. *Nature* **366**, 557–559 (1993).
2. DeCelles, P. G. *et al.* Stratigraphy, structure, and tectonic evolution of the Himalayan fold-thrust belt in western Nepal. *Tectonics* **20**, 487–509 (2001).
3. Bollinger, L. *et al.* Thermal structure and exhumation history of the Lesser Himalaya in Central Nepal. *Tectonics* **23**, doi:10.1029/2003TC001564 (2004).
4. Dueker, K. G. & Sheehan, A. F. Mantle discontinuity structure beneath the Colorado Rocky Mountains and High Plains. *J. Geophys. Res.* **103**, 7153–7169 (1998).
5. Kind, R. *et al.* Seismic images of crust and upper mantle beneath Tibet: evidence for Eurasian plate subduction. *Science* **298**, 1219–1221 (2002).
6. Brunel, M. Ductile thrusting in the Himalayas; shear sense criteria and stretching lineations. *Tectonics* **5**, 247–265 (1986).
7. Schelling, D. The tectonostratigraphy and structure of the eastern Nepal Himalaya. *Tectonics* **11**, 925–943 (1992).
8. Avouac, J.-P. Mountain building, erosion, and the seismic cycle in the Nepal Himalaya. *Adv. Geophys.* **46**, 11–80 (2003).
9. Zandt, G., Gilbert, H., Owen, T., Ducea, M. & Jones, C. Active foundering of a

- continental arc root beneath the southern Sierra Nevada in California. *Nature* **431**, 41–44 (2004).
10. Bilham, R., Larson, K. & Freymüller, J. GPS measurements of present-day convergence across the Nepal Himalaya. *Nature* **386**, 61–64 (1997).
11. Daniel, C., Hollister, L., Parrish, R. & Grujic, D. Exhumation of the Main Central Thrust from lower crustal depths, eastern Bhutan Himalaya. *J. Metamorph. Geol.* **21**, 317–334 (2003).
12. Kumar, M. R., Saul, J., Sarkar, D., Kind, R. & Shukla, A. Crustal structure of the Indian shield; new constraints from teleseismic receiver functions. *Geophys. Res. Lett.* **28**, 1339–1342 (2001).
13. Le Pichon, X., Fournier, M. & Jolivet, L. Kinematics, topography, shortening, and extrusion in the India-Asia collision. *Tectonics* **11**, 1085–1098 (1992).
14. Monsalve, G. *et al.* Earthquake locations, depths and source parameters from the Himalaya Nepal Tibet seismic experiment: Implications for continental lithospheric depth. *Eos* **84**(46), S22A–0414 (2003).
15. Pandey, M. R., Tandukar, R. P., Avouac, J.-P., Lave, J. & Massot, J.-P. Interseismic strain accumulation on the Himalayan crustal ramp (Nepal). *Geophys. Res. Lett.* **22**, 751–754 (1995).
16. Christensen, N. & Mooney, W. Seismic velocity structure and composition of the continental crust: A global view. *J. Geophys. Res.* **100**, 9761–9788 (1995).
17. Bousquet, R., Goffé, B., Henry, P., Le Pichon, X. & Chopin, C. Kinematic, thermal and petrological model of the central Alps: Lepontine metamorphism in the upper crust and eclogitization of the lower crust. *Tectonophysics* **273**, 105–127 (1997).
18. Bjørnerud, M., Austrheim, H. & Lund, M. G. Processes leading to eclogitization (densification) of subducted and tectonically buried crust. *J. Geophys. Res.* **107**, doi:10.1029/2001JB000527 (2002).
19. Lyon-Caen, H. & Molnar, P. Gravity anomalies, flexure of the Indian plate, and the structure, support and evolution of the Himalaya and Ganga basin. *Tectonics* **4**, 513–538 (1985).
20. Cattin, R. *et al.* Gravity anomalies, crustal structure and thermo-mechanical support of the Himalaya of central Nepal. *Geophys. J. Int.* **147**, 381–392 (2001).
21. Hirn, A. & Sapin, M. The Himalayan zone of crustal interaction; suggestions from explosion seismology. *Ann. Geophys.* **2**, 123–130 (1984).
22. Molnar, P. A review of geophysical constraints on the deep structure of the Tibetan Plateau, the Himalaya and the Karakoram, and their geophysical implications. *Phil. Trans. R. Soc. Lond A* **326**, 33–88 (1988).
23. Galvé, A. *et al.* Complex images of Moho and variation of Vp/Vs across the Himalaya and South Tibet, from a joint receiver-function and wide-angle-reflection approach. *Geophys. Res. Lett.* **29**, doi:10.1029/2002GL015611 (2002).
24. Sapin, M. & Hirn, A. Seismic structure and evidence for eclogitization during the Himalayan convergence. *Tectonophysics* **273**, 1–16 (1997).
25. Johnson, M. R. W. Shortening budgets and the role of continental subduction during the India-Asia collision. *Earth Sci. Rev.* **59**, 101–123 (2002).
26. Cotte, N. *et al.* Determination of the crustal structure in southern Tibet by dispersion and amplitude analysis of Rayleigh waves. *Geophys. J. Int.* **138**, 809–819 (1999).
27. Zhao, W., Sen, M., Stoffa, P. & Frohlich, C. Application of very fast simulated annealing to the determination of the crustal structure beneath Tibet. *Geophys. J. Int.* **125**, 355–370 (1996).
28. Jackson, J. A., Austrheim, H., McKenzie, D. & Priestley, K. Metastability, mechanical strength, and the support of mountain belts. *Geology* **32**, 625–628 (2004).
29. Chen, W. & Yang, Z. Earthquakes beneath the Himalaya and Tibet: Evidence for a strong lithospheric mantle. *Science* **304**, 1949–1952 (2004).
30. Henry, P., Le Pichon, X. & Goffé, B. Kinematic, thermal and petrological model of the Himalayas; constraints related to metamorphism within the underthrust Indian crust and topographic elevation. *Tectonophysics* **273**, 31–56 (1997).
31. Le Pichon, X., Henry, P. & Goffé, B. Uplift of Tibet; from eclogites to granulites; implications for the Andean plateau and the Variscan belt. *Tectonophysics* **273**, 57–76 (1997).
32. Frederiksen, A. & Bostock, M. Modelling teleseismic waves in dipping anisotropic structures. *Geophys. J. Int.* **141**, 401–412 (2000).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank C. Jones and P. Molnar for comments; H. Gilbert, F. Blume, R. Bendick, the Department of Mines and Geology of Nepal, particularly R. K. Aryal and N. R. Shtapit, and the Chinese Academy of Sciences, in particular H. B. Liu, for their collaboration and assistance; the US National Science Foundation for support; and the Incorporated Research Institutions for Seismology for use of their instruments.

Author Contributions V.S., A.S., S.S., M.-P., R.B. and F.W. participated in the deployment and data collection. V.S., G.M. and A.S. performed the seismological analysis. S.S., R.B. and F.W. provided the tectonic and geological context. V.S. wrote the Letter with contributions from all authors.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to V.S. (vera.schulte-pelkum@colorado.edu).

LETTERS

Experimental demonstration of chaos in a microbial food web

Lutz Becks^{1*}, Frank M. Hilker², Horst Malchow², Klaus Jürgens^{3,4} & Hartmut Arndt^{1*}

Discovering why natural population densities change over time and vary with location is a central goal of ecological and evolutionary disciplines. The recognition that even simple ecological systems can undergo chaotic behaviour has made chaos a topic of considerable interest among theoretical ecologists^{1–4}. However, there is still a lack of experimental evidence that chaotic behaviour occurs in the real world of coexisting populations in multi-species systems. Here we study the dynamics of a defined predator–prey system consisting of a bacterivorous ciliate and two bacterial prey species. The bacterial species preferred by the ciliate was the superior competitor. Experimental conditions were kept constant with continuous cultivation in a one-stage chemostat. We show that the dynamic behaviour of such a two-prey, one-predator system includes chaotic behaviour, as well as stable limit cycles and coexistence at equilibrium. Changes in the population dynamics were triggered by changes in the dilution rates of the chemostat. The observed dynamics were verified by estimating the corresponding Lyapunov exponents. Such a defined microbial food web offers a new possibility for the experimental study of deterministic chaos in real biological systems.

Apart from the intuitive understanding that external (extrinsic) stimuli influence the variability of abundances, mathematical models have made it apparent that the internal (intrinsic) qualities of a population give rise to population dynamics with large and (at certain parameter ranges) even chaotic fluctuations of abundances, even under wholly constant and predictable conditions⁵. Predator–prey interactions have been considered as a possible driving force of population dynamics since the beginning of ecological studies^{6,7}. In his analysis of mathematical models, May¹ found that even simple processes of population growth can show (for a certain range of parameters) an unpredictable behaviour driven by intrinsic mechanisms. May's studies marked the beginning of an intensive debate on the question of whether or not natural systems are characterized by chaotic behaviour. In this context, the term 'deterministic chaos' can be defined as bounded aperiodic fluctuations with sensitive dependence on initial conditions⁴. Under chaotic conditions, population abundances never show a precisely repeated pattern over time; such patterns are only observable in populations at equilibrium or at stable limit cycles. Theoreticians can clearly define parameter ranges of mathematical models that create chaotic behaviour in idealized biological systems^{3,8–10}. However, only a very few experiments indicating that bifurcations of dynamic behaviour might occur in the real world have been conducted (for example, ciliate–bacteria interactions¹¹, flour beetle (*Tribolium castaneum*) dynamics^{12,13} and rotifer–algae interactions¹⁴). Indications of chaotic dynamics under controlled conditions have so far been reported for one-species systems only¹³. A robust tool to verify observed dynamics is

estimations of Lyapunov exponents from time series, which test for the exponential divergence of nearby trajectories. Mathematically, stable (convergent) systems show negative Lyapunov exponents, whereas chaotic (divergent) systems have at least one positive Lyapunov exponent⁴.

The aim of the present study was to verify the biological relevance of chaotic behaviour in a real multi-species system. The long generation durations of most organisms and the complexity of natural environments have generally made the explanation of underlying ecological mechanisms difficult¹⁵. However, experiments using microbial populations propagated in controlled environments reduce ecosystem complexity to the point at which understanding simple processes in isolation becomes possible. The rapid reproduction of bacteria and protists is one of the main advantages of working with microorganisms as model organisms^{7,16,17}. In addition, the community structure can be exactly defined; for example, single strains of bacteria and protists can be selected. Microorganisms can also be cultured under chemostat conditions. This has the great advantage that extrinsic factors are negligible and changes in population dynamics can be attributed to intrinsic factors. In terms of predation and interspecific competition, one of the simplest systems imaginable is a three-species system with two prey organisms and one predator. Several theoretical studies have been made of such model systems^{8–10,18}. Generally, different patterns of population dynamics are predicted by models; for example, the extinction of one or two species and the coexistence of all three species. Assuming that the two prey populations compete with each other and assuming that the better competitor is the preferred prey, three patterns may occur: coexistence at equilibrium, coexistence at stable limit cycles, and coexistence at chaos^{8–10,18}.

Our study was aimed at identifying these different patterns of coexistence in controlled experiments in a chemostat. We used the dilution rate as the bifurcation parameter in the experiments, because the dynamical behaviour of chemostat models can change with dilution rate^{9,10,14}. We constructed one-stage chemostat systems consisting of axenic cultures of three species: a predator (the ciliate *Tetrahymena pyriformis*) and two coexisting prey bacteria, the rod-shaped *Pedobacter* and the coccus *Brevundimonas*. The effective consumption of these bacteria by the ciliate and its food preference was analysed by immunofluorescence techniques. The ciliate can establish stable populations when feeding on either bacterium, but it dies off in the highly diluted organic medium when bacteria are absent. The growth conditions for the bacteria and the mortality of the ciliate are determined by the dilution rates (controlled by peristaltic pumps). *Brevundimonas* was always outcompeted in chemostat experiments containing both bacterial strains without a predator. Thus, *Pedobacter* was considered to have a better fitness. In

¹Department of General Ecology and Limnology, Zoological Institute, University of Cologne, D-50923 Köln, Germany. ²Department of Mathematics and Computer Science, Institute of Environmental Systems Research, University of Osnabrück, D-49069 Osnabrück, Germany. ³Max Planck Institute for Limnology, PO Box 165, D-24302 Plön, Germany. ⁴Baltic Sea Research Institute, D-18119 Rostock-Warnemünde, Germany.

*These authors contributed equally to this work.

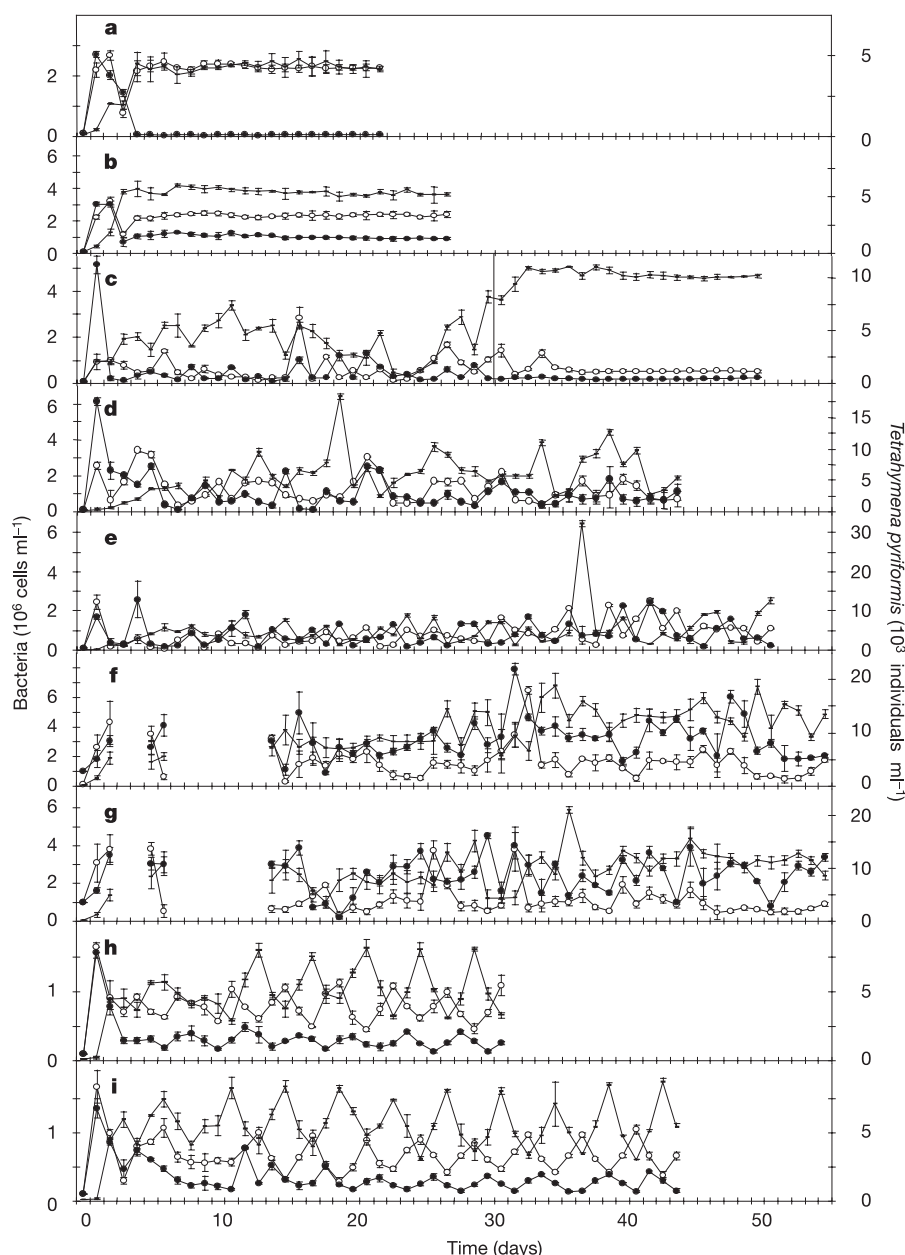


Figure 1 | Experimental results showing the population dynamics of bacteria-ciliate chemostat systems. Dilution rates D were as follows: **a**, 0.90 d^{-1} ; **b**, 0.75 d^{-1} ; **c**, 0.50 d^{-1} (the line indicates the change to 0.75 d^{-1} at day 30); **d–g**, 0.50 d^{-1} (replicate experiments; no sampling took place on

days 3, 4 and 7–13 in **f** and **g**); **h**, **i**, 0.45 d^{-1} (replicate experiments). Open circles, *Pedobacter* (preferred prey); filled circles, *Brevundimonas* (less-preferred prey); horizontal bar, *Tetrahymena* (predator). Vertical bars represent the s.d. of triplicate samples taken separately from one chemostat.

contrast, our grazing experiments revealed that *Pedobacter* is preferred as prey by the ciliate over *Brevundimonas* by a factor of four (see Supplementary Information). Experiments were performed with dilution rates of 0.90, 0.75, 0.50 and 0.45 d^{-1} . These dilution rates were selected on the basis of preceding model calculations.

The results revealed a different dynamic behaviour of the experimental system, depending on the applied dilution rate (Fig. 1). At the highest dilution rate ($D = 0.90 \text{ d}^{-1}$), *Brevundimonas* had died off by the sixth day; the remaining species existed in stable coexistence at equilibrium (Fig. 1a). The establishment of constant, equilibrium population densities of all species was achieved after about 5 days at $D = 0.75 \text{ d}^{-1}$ (Fig. 1b). To check the robustness of the stable equilibrium, we repeated this experiment with a preceding 30-day period of aperiodic dynamics at $D = 0.50 \text{ d}^{-1}$ (Fig. 1c). Although similar dynamic behaviour was observed after a transition period of 5 days, the abundances reached by the three species were different from

those of the previous experiment. One possible explanation for these differences in abundances of the rapidly reproducing microbes (30 days represents about 240 generations in the experiments) might be a potential evolutionary shift in population structure¹⁹. Obviously stable limit cycles were established in the two parallel chemostat systems after a period of about 8 days at a dilution rate of 0.45 d^{-1} (Fig. 1h, i). Maxima and minima for all three species recurred during the whole observation period. Slight differences can be attributed to the sampling interval, which was kept constant at about 24 h. The cycles started with a maximum abundance of the preferred bacterium, followed by a peak of the less-preferred bacterium and the predator. Aperiodic oscillations were always obtained when dilution rates were set to 0.50 d^{-1} (Fig. 1d–g). All four trials showed different patterns in their dynamics. The observed aperiodic oscillations of the chemostat populations were analysed for possible chaotic behaviour by using estimates of corresponding Lyapunov exponents (Fig. 2; see

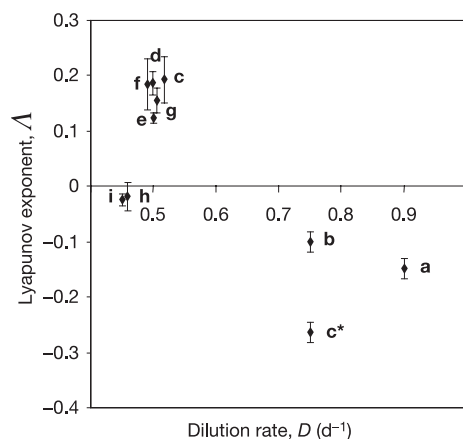


Figure 2 | Relationship between trajectory stability of data sets from Fig. 1 and the corresponding dilution rates. Lyapunov exponents were determined by the method of Rosenstein *et al.*²⁰ by fitting the rate of exponential separation of initially close trajectories. The error bars correspond to the asymptotic errors in the fit. Letters correspond to panels in Fig. 1 (**c*** is the Lyapunov exponent calculated for the second part of the time series in Fig. 1c; $t = 31$ –50 days). Note that for dilution rates of 0.45 d^{-1} and 0.5 d^{-1} , data points were spread slightly along the x axis for visual clarity.

Supplementary Information). They were determined from each time series with a previously published algorithm²⁰, which tests directly for the exponential divergence from nearby trajectories and provides a very robust method for also dealing with small data sets (see Methods). According to general theoretical expectations, the data sets with extinction of the less-preferred prey species (Fig. 1a) and with coexistence at stable equilibrium (Fig. 1b) revealed negative Lyapunov exponents. This also holds true for the second part of the time series in Fig. 1c (after a change in the dilution rate to 0.75 d^{-1}). All experiments with $D = 0.5 \text{ d}^{-1}$ (Fig. 1c, left of the vertical line, and Fig. 1d–g) have positive Lyapunov exponents. Thus, we obtained strong experimental evidence for the existence of chaos in a real multi-species system. Note that for small data sets the range of the confidence interval generally increases. The stable sustained oscillations (Fig. 1h, i) have Lyapunov exponents close to zero (Fig. 2). Their absolute value is at least one order of magnitude smaller than all the other exponents. The exponential divergences of nearby trajectories show strong, sustained oscillations as well. There is a large asymptotic standard error in the fit of the Lyapunov exponents because of the strong oscillations. We conclude that the underlying dynamics are stable limit cycles. The observed dynamics in the experiments changed as predicted by a model¹⁰ (stable coexistence at high dilution rates, chaos at intermediate dilution rates, and stable limit cycles at low dilution rates).

There are two important conclusions to be drawn. First, chaotic dynamics of small, rapidly growing organisms can occur in all micro-biotopes. Second, because of the low generation times of microbes (only a few hours), such dynamics may be established before perturbations by external stimuli are effective. Examples of such communities are the tiny, fragmented populations of protists and bacteria that can occur on each grain of sand in a sediment, as well as on each small detritus particle in the pelagic zone of the open ocean or lakes^{21,22}. The defined microbial food web that we established under chemostat conditions offers a completely new possibility for the experimental study of deterministic chaos in real biological systems. It is now possible to address many questions previously posed by theoreticians. We have provided a biological system that allows the investigation of the transition between different dynamical states, the analysis of interactions of fragmented populations showing either similar or different dynamic behaviours, the study of

resilience and the importance of perturbations under varying dynamical states, and the interplay between complex dynamics and biodiversity^{1–4,8,16,18,23}. When combined with molecular techniques, this system would also allow the evolutionary consequences of different dynamic behaviours to be analysed¹⁹.

METHODS

Chemostat experiments. We established cultures of the ciliate *Tetrahymena pyriformis* (axenic culture from CCAP 1630/1W, average length and width $85 \mu\text{m} \times 22 \mu\text{m}$), the bacterium *Pedobacter* sp. (Cytophaga Flexibacter group, $2 \mu\text{m} \times 1 \mu\text{m}$) and *Brevundimonas* sp. (α -Proteobacteria, $2.5 \mu\text{m} \times 2.5 \mu\text{m}$) in 185 ml glass chemostats at $20 \pm 1^\circ\text{C}$ in the dark. Both bacterial species were isolated by K. Beck from Lake Schöhsee, Germany; bacteria were always inoculated from deep-frozen stock cultures. The one-stage chemostat systems were fed continuously with sterile medium (0.2 g l^{-1} proteose peptone, 0.025 g l^{-1} yeast extract) at different dilution rates and mixed by continuous gentle aeration to ensure an even distribution of organisms. Chemostats were always started with the same inoculum. Sterile syringes were used to take samples daily at about 11:00 from the centre of the chemostats. Living ciliate samples were counted under a phase-contrast microscope immediately after sampling (more than 150 individuals were counted). Samples of bacteria were fixed with formaldehyde and stained with 4',6-diamidino-2-phenylindole (DAPI)²⁴ for subsequent counting on membrane filters (pore size $0.2 \mu\text{m}$) under an epifluorescence microscope (Zeiss Axioskop) with Zeiss filter set 01. At least 300 bacteria were counted on each filter. Organism abundances were the average of triplicates taken separately from one chemostat. The total volume of water taken from the chemostats during one sampling was 3 ml. Chemostats were checked regularly for the appearance of contaminant bacteria by using strain-specific antibodies against *Pedobacter* and *Brevundimonas* and by non-specific staining of the bacterial community with DAPI. With our present apparatus, the maximum number of samplings possible before contamination or any other technical problem hindered further experimentation was 50–55 days.

Grazing experiments. Experiments were performed to determine the food preference of *Tetrahymena*. A bacterial mixture (1:1) of *Pedobacter* and *Brevundimonas* (each strain at $4 \times 10^6 \text{ cells ml}^{-1}$) was offered as prey in 50-ml vessels at 20°C . The contents of the vessels were fixed with a buffered paraformaldehyde solution 3 min after inoculation of *Tetrahymena*²⁵. The abundances of *Pedobacter* and *Brevundimonas* in the food vacuoles of *Tetrahymena* were determined by immunofluorescence²⁶ after hybridization with specific Cy3-labelled antibodies (permeabilization was performed with 8% Triton X-100).

Calculation of Lyapunov exponents. The calculations of the Lyapunov exponents by using the algorithm of Rosenstein *et al.*²⁰ were performed with the TISEAN package²⁷ (see Supplementary Information). Similarly to the independently published algorithm of Kantz²⁸, it directly tests the presence of exponential divergence and thus permits a decision on whether it makes sense to compute a Lyapunov exponent for given data. In contrast, the first published and widely used algorithm of Wolf *et al.*²⁹ makes the *a priori* assumption that there is an exponential divergence of nearby trajectories and is therefore prone to yield finite positive Lyapunov exponents also for stochastic data. This has been criticized in the ecological literature^{4,30}, and alternative approaches have been proposed that rely on approximating the equations of the underlying dynamics. The exponents are calculated from the jacobian, which resembles the linear part of the dynamics. This method is efficient if the data permit a good reconstruction of the dynamics. However, one has to be careful, because a good approximation of the dynamics does not guarantee well-approximated partial derivatives in the jacobian. However, because the present data stem from constant experimental conditions in a chemostat environment, the algorithm of Rosenstein *et al.*²⁸ should reveal more reliable estimates. The exponents were calculated by reconstructing the attractor dynamics from the time series of the predator's abundances with appropriate embedding dimensions and reconstruction delays, which robustly exhibited exponential divergence. The Lyapunov exponent was then fitted as the slope of the linear increase in the log-transformed divergence by using the least-squares method.

Received 2 March; accepted 12 April 2005.

- May, R. M. Biological populations with nonoverlapping generations: stable points, stable cycles, and chaos. *Science* **186**, 645–647 (1974).
- Hastings, A. & Powell, T. Chaos in a three-species food chain. *Ecology* **72**, 896–903 (1991).
- Huisman, J. & Weissing, F. J. Biodiversity of plankton by species oscillations and chaos. *Nature* **402**, 407–411 (1999).
- Turchin, P. *Complex Population Dynamics: A Theoretical/Empirical Synthesis* (Princeton Univ. Press, Princeton, 2003).

5. Begon, M., Harper, J. L. & Townsend, C. R. *Ecology* (Blackwell Science, Oxford, 2002).
6. Volterra, V. Fluctuations in the abundance of a species considered mathematically. *Nature* **118**, 558–600 (1926).
7. Luckenbill, L. S. & Fenton, M. M. Regulation and environmental variability in experimental populations of protozoa. *Ecology* **59**, 1271–1276 (1978).
8. Takeuchi, Y. & Adachi, N. Existence and bifurcation of stable equilibrium in two-prey-one-predator communities. *Bull. Math. Biol.* **45**, 877–900 (1983).
9. Vayenas, D. V. & Pavlou, S. Chaotic dynamics of a food web in a chemostat. *Math. Biosci.* **162**, 69–84 (1999).
10. Kooi, B. W. & Boer, M. P. Chaotic behaviour of a predator–prey system in the chemostat. *Dynam. Contin. Discrete Impulsive Syst. B* **10**, 259–272 (2003).
11. Jost, J. L., Drake, J. F., Frederickson, A. E. & Tsuchiya, H. M. Interactions of *Tetrahymena pyriformis*, *Escherichia coli*, *Azotobacter vinelandii*, and glucose in a mineral medium. *J. Bacteriol.* **113**, 834–840 (1973).
12. Costantino, R. F., Desharnais, R. A., Cushing, J. M. & Dennis, B. Chaotic dynamics in an insect population. *Science* **275**, 389–391 (1997).
13. Dennis, B., Desharnais, R. A., Cushing, J. M. & Costantino, R. F. Transitions in population dynamics: equilibria to periodic cycles to aperiodic cycles. *J. Anim. Ecol.* **66**, 704–729 (1997).
14. Fussmann, G. F., Ellner, S. P., Shertzer, K. W. & Hairston, N. G. Jr Crossing the Hopf bifurcation in a live predator–prey system. *Science* **290**, 1358–1360 (2000).
15. Heerkloss, R. & Klinkenberg, G. Chaotic dynamics of a plankton community in a species-depleted mesocosmos. *Verh. Int. Verein. Limnol.* **25**, 995–1000 (1993).
16. Rainey, P. B., Buckling, A., Kassen, R. & Travisano, M. The emergence and maintenance of diversity: insights from experimental bacterial populations. *Trends Ecol. Evol.* **15**, 243–247 (2000).
17. Jessup, C. M. et al. Big questions, small worlds: microbial model systems in ecology. *Trends Ecol. Evol.* **19**, 189–197 (2004).
18. Gilpin, M. E. Spiral chaos in a predator–prey model. *Am. Nat.* **113**, 306–308 (1979).
19. Yoshida, T. et al. Rapid evolution drives ecological dynamics in a predator–prey system. *Nature* **424**, 303–306 (2003).
20. Rosenstein, M. T., Collins, J. J. & De Luca, C. J. A practical method for calculating largest Lyapunov exponent from small data sets. *Physica D* **65**, 117–134 (1993).
21. Arndt, H. et al. in *The Flagellates* (eds Leadbeater, B. S. C. & Green, J. C.) 240–268 (Taylor & Francis, London, 2000).
22. Caron, D., Davis, P. G., Madin, L. P. & Sieburth, J. M. Heterotrophic bacteria and bacterivorous protozoa in oceanic macroaggregates. *Science* **218**, 795–797 (1982).
23. Nelson, W. A., McCauley, E. & Wrona, F. J. Stage-structured cycles promote genetic diversity in a predator–prey system of *Daphnia* and algae. *Nature* **433**, 413–417 (2005).
24. Porter, K. G. & Feig, Y. S. The use of DAPI for identifying and counting aquatic microflora. *Limnol. Oceanogr.* **25**, 943–948 (1980).
25. Eisenmann, H., Harms, H., Meckenstock, R., Meyer, E. I. & Zehnder, A. J. B. Grazing of a *Tetrahymena* sp. on adhered bacteria in percolated columns monitored by in situ hybridization with fluorescent oligonucleotide probes. *Appl. Environ. Microbiol.* **64**, 1264–1269 (1998).
26. Christofferson, K., Nybroe, O., Jürgens, K. & Hansen, M. Measurement of bacterivory by heterotrophic nanoflagellates using immunofluorescence labelling of ingested cells. *Aquat. Microb. Ecol.* **13**, 127–134 (1997).
27. Hegger, R., Kantz, H. & Schreiber, T. Practical implementation of nonlinear time series methods: The TISEAN package. *Chaos* **9**, 413–439 (1999).
28. Kantz, H. A robust method to estimate the maximal Lyapunov exponent of a time series. *Phys. Lett. A* **185**, 77–87 (1994).
29. Wolf, A., Swift, J. B., Swinney, H. L. & Vastano, J. A. Determining Lyapunov exponents from a time series. *Physica D* **16**, 285–317 (1985).
30. Ellner, S. P. & Turchin, P. Chaos in a noisy world: New methods and evidence from time-series analysis. *Am. Nat.* **145**, 343–375 (1995).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank J. Huisman, M. Milinski, A. Scherwass, D. Stauffer, D. Tautz and M. Weitere for their constructive comments on the manuscript; R. Heerkloss for introducing us to the topic; W. Lampert for stimulating atmosphere during sabbatical in Plön; F. Bartlett for correction of the English manuscript; and C. Barth, R. Bieg and B. Gräfe for technical assistance. The study was supported by a grant from the German Research Foundation (Deutsche Forschungsgemeinschaft) to H.A.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to H.A. (hartmut.arndt@uni-koeln.de).

LETTERS

Clonal reproduction by males and females in the little fire ant

Denis Fournier^{1*†}, Arnaud Estoup^{1*}, Jérôme Orivel², Julien Foucaud¹, Hervé Jourdan³, Julien Le Breton⁴ & Laurent Keller⁵

Sexual reproduction can lead to major conflicts between sexes and within genomes^{1–4}. Here we report an extreme case of such conflicts in the little fire ant *Wasmannia auropunctata*. We found that sterile workers are produced by normal sexual reproduction, whereas daughter queens are invariably clonally produced. Because males usually develop from unfertilized maternal eggs in ants and other haplodiploid species, they normally achieve direct fitness only through diploid female offspring. Hence, although the clonal production of queens increases the queen's relatedness to reproductive daughters, it potentially reduces male reproductive success to zero. In an apparent response to this conflict between sexes, genetic analyses reveal that males reproduce clonally, most likely by eliminating the maternal half of the genome in diploid eggs. As a result, all sons have nuclear genomes identical to those of their father. The obligate clonal production of males and queens from individuals of the same sex effectively results in a complete separation of the male and female gene pools. These findings show that the haplodiploid sex-determination system provides grounds for the evolution of extraordinary genetic systems and new types of sexual conflict.

The little fire ant has been introduced from neotropical lowland forests into North America, West Africa, Melanesia, Polynesia, the Galapagos and some subtropical Atlantic Islands, where it has become a major pest. Colonies consist of several spatially separated nests headed by multiple reproductive queens⁵. Although queens can participate in mating flights⁶, colonies spread largely or entirely by budding, a process in which one or more queens initiate a new colony in the vicinity of the mother nest with the help of workers⁵. While conducting a genetic population study of this species, we discovered a new genetic system in which females and males both reproduce clonally.

We collected 34 nests of *W. auropunctata* from five sites in French Guiana, which is within the native range (Fig. 1). The number of queens per nest was 4.2 ± 0.7 (mean $\pm$ s.e.m.; range 0–18). We genotyped at 11 highly polymorphic microsatellite loci (observed heterozygosities: 0.502–0.964) all the queens ($n = 142$) collected, the sperm in their spermathecae, the nine young winged queens (gynes) found in one of the 34 nests, and 264 workers (7.8 ± 0.2 workers per nest). An analysis of these genotypes revealed a very unusual pattern that could be explained only by queens being produced by clonal

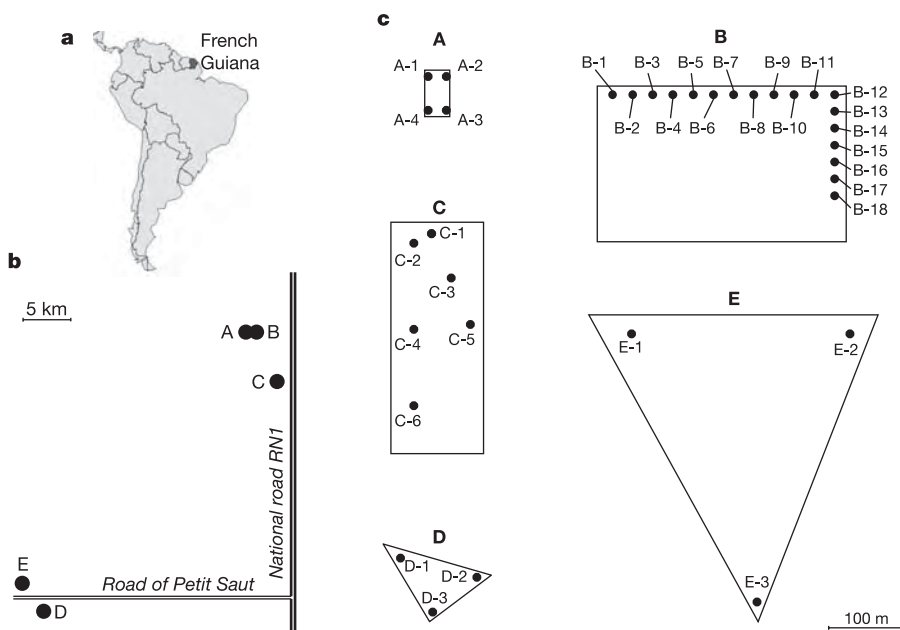


Figure 1 | Site and nest locations. **a, b**, Locations of the five sites of collection (A–E): national (**a**) and local (**b**). **c**, Locations of the nests within each of the five sites; scales are identical for each of the five sites.

¹Centre de Biologie et de Gestion des Populations, INRA, Campus International de Baillarguet, CS 30 016, 34988 Montpellier/Lez Cedex, France. ²Laboratoire Évolution et Diversité Biologique, UMR-CNRS 5174, Université Toulouse III, 118 route de Narbonne, 31062 Toulouse cedex 4, France. ³Laboratoire de Zoologie Appliquée, IRD, 98948 Nouméa, Nouvelle-Calédonie. ⁴Laboratory of Sub-Tropical Zoology, University of the Ryukyus, 903-0213 Nishihara, Okinawa, Japan. ⁵Department of Ecology and Evolution, Bâtiment de Biologie, University of Lausanne, 1015 Lausanne, Switzerland. [†]Present address: Behavioral and Evolutionary Ecology – CP 160/12, Université Libre de Bruxelles, 50 avenue F. D. Roosevelt, 1050 Brussels, Belgium.

*These authors contributed equally to this work.

Table 1 | Genotypes of queens (Q), their mates (M) and workers (w) in one nest (E-3) at each of the 11 microsatellite loci

Individual	Waur-225		Waur-275		Waur-418		Waur-566		Waur-680		Waur-716		Waur-730		Waur-1166		Waur-2164		Waur-3176		Waur-1gam	
Queens																						
Q-1	223	225	105	115	100	112	263	263	171	171	184	198	158	160	95	97	298	306	230	230	288	298
Q-2	223	225	105	115	100	112	263	263	171	171	184	198	158	160	95	97	298	306	230	230	288	298
Q-3	223	225	105	115	100	112	263	263	171	171	184	198	158	160	95	97	298	306	230	230	288	298
Q-4	223	225	105	115	100	112	263	263	171	171	184	198	158	160	95	97	298	306	230	230	288	298
Q-5	223	225	105	115	100	112	263	263	171	171	184	198	158	160	95	97	298	306	230	230	288	298
Q-6	223	225	105	115	100	112	263	263	171	171	184	198	158	160	95	97	298	306	230	230	288	298
Q-7	223	225	105	115	100	112	263	263	171	171	184	198	158	160	95	97	298	306	230	230	288	298
Q-8	223	225	105	115	100	112	263	263	171	171	184	198	158	160	95	97	298	306	230	230	288	298
Males																						
M-1		269		107		118		265		187		192		214		95		320		244		282
M-2		269		107		118		265		187		192		214		95		320		244		282
M-3		269		107		118		265		187		192		214		95		320		244		282
M-4		269		107		118		265		187		192		214		95		320		244		282
M-5		269		107		118		265		187		192		214		95		320		244		282
M-6		269		107		118		265		187		192		214		95		320		244		282
M-7		269		107		118		265		187		192		214		95		320		244		282
M-8		269		107		118		265		187		192		214		95		320		244		282
Workers																						
w-1	223	269	115	107	112	118	263	265	171	187	198	192	160	214	95	95	306	320	230	244	298	282
w-2	225	269	115	107	100	118	263	265	171	187	184	192	158	214	95	95	298	320	230	244	288	282
w-3	223	269	105	107	112	118	263	265	171	187	198	192	160	214	97	95	298	320	230	244	298	282
w-4	225	269	115	107	100	118	263	265	171	187	184	192	158	214	97	95	306	320	230	244	288	282
w-5	223	269	105	107	100	118	263	265	171	187	198	192	158	214	97	95	306	320	230	244	298	282
w-6	225	269	115	107	112	118	263	265	171	187	184	192	160	214	97	95	306	320	230	244	288	282
w-7	223	269	105	107	100	118	263	265	171	187	184	192	158	214	97	95	306	320	230	244	298	282
w-8	225	269	115	107	112	118	263	265	171	187	184	192	158	214	97	95	298	320	230	244	288	282

The identities of mates were determined by the sperm collected in the queen's spermathecae. Queens and males' genotypes illustrate their clonal production, whereas workers' genotypes are consistent with normal sexual reproduction. Paternal alleles are in *italics*.

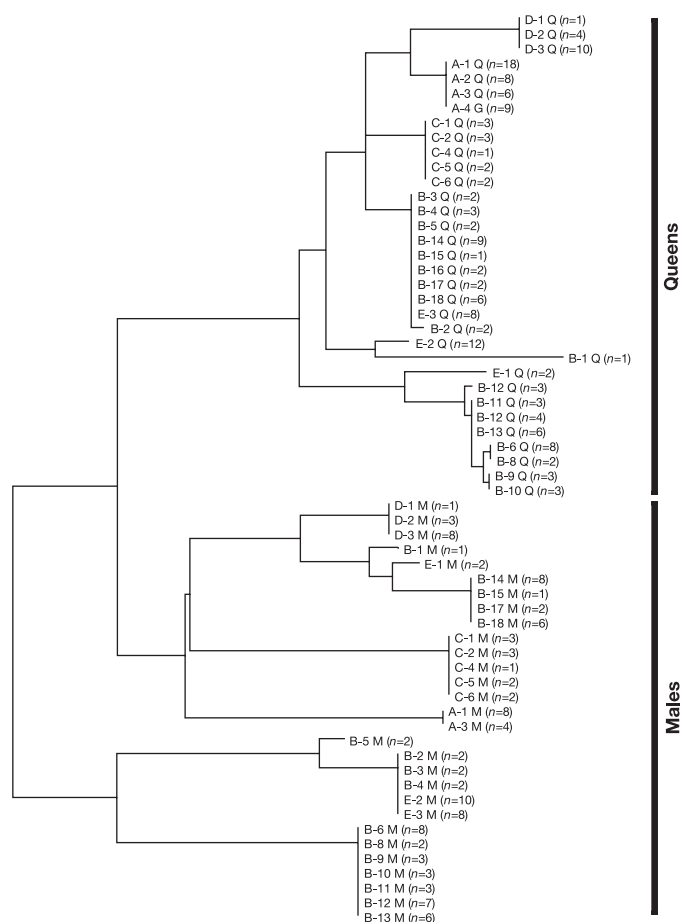


Figure 2 | Neighbour-joining dendrogram of the genetic (allele-shared) distances between queens (Q), gynes (G) and male sperms (M) collected over all the five sites (A-E). The collection number of each nest is given with the letter of the site (see Fig. 1 for details). The number of individuals sharing the same genotype (*n*) is given for all nests.

reproduction (that is, by ameiotic parthenogenesis). In 33 of the 34 nests, all queens ($n = 135$) and gynes ($n = 9$) cohabiting in the same nest shared an identical genotype at each of the 11 loci (Table 1 and Fig. 1). The single exception was nest B-12, in which queens differed at 1 of the 11 loci: four queens were heterozygous at *Waur-2164* and the remaining three queens were homozygous for one of the two alleles. This variation probably reflects a mutation or recombination event in one queen followed by clonal reproduction within the nest. The history of this genetic change could be reconstructed from the genotypes of queens collected in neighbouring nests (Figs 1 and 2). Nine queens from two neighbouring nests (B-11 and B-13) had the same genotype as the four heterozygous queens for locus *Waur-2164*, indicating that the mutation or recombination event probably was from a heterozygote to a homozygote queen. The three homozygote queens from nest B-12 had a unique genotype in the population, which further supports this interpretation.

A comparison between nests supports the view of restricted female gene flow, with budding being the main mode of colony formation. Within three of the five sites of collection (A, C and D) all queens had the same genotype at the 11 loci (Fig. 2). In one of the two other sites (B), all queens from 8 of the 17 nests also had an identical genotype, whereas in the other site (E) the queen genotypes were different in the three nests sampled. Taken together, these data indicate that queens belonging to the same lineage of clonally produced individuals frequently head closely located nests. Moreover, genetic differentiation between sites was very strong, with a single occurrence of genotypes shared between sites (the eight queens of nest E-3 had genotypes identical to the most common genotype found at site B), showing that gene flow by females is extremely restricted.

In stark contrast to reproductive females, the genotypic analyses revealed that workers are produced by normal sexual reproduction (Table 1). Over all 31 queenright nests, each of the 248 genotyped workers had, at seven or more loci, one allele that was absent in queens of their nest. Moreover, the 232 workers from the 29 nests in which the sperm in the queen's spermathecae was successfully obtained had all genotypes consistent with those expected under sexual reproduction between the two parental genomes.

The genetic analyses of the sperm collected in the queens'

spermathecae revealed a pattern of partitioning of genetic variation that was remarkably similar to the one found in reproductive females. Cohabiting queens from a given nest were always inseminated by males having an identical genotype at the 11 loci (Table 1). Moreover, a single male genotype was found in the three sites (A, C and D) harbouring a single queen genotype. In contrast, the two remaining sites (B and E) that contained several queen genotypes also had several male genotypes. The observation that nests never contained more than one male genotype also supports the view that male dispersal is also extremely limited, with most or all matings taking place within the parental nest. This breeding system would account for nests invariably containing a unique queen and male genotype.

The finding that the males that inseminated queens of a given nest invariably had the same genotype, together with the fact that males and queens never have the same genotype, reveals that the genomes of males are also transmitted clonally. Additional evidence that males are indeed clonally produced from the sperm in the queens' spermathecae came from the genotypes of 41 males found in four of the nests (mean $\pm$ s.e.m. 10.3 ± 2.5) collected in another population in New Caledonia. In each of the four colonies, males had genotypes (11 microsatellites) identical to those of the sperm found in the spermathecae of queens heading their colony (the 12.0 ± 1.5 queens per colony again had all the same genotype) and genotypes incompatible with maternal inheritance of their genomes (all males had alleles that were absent in queens at 10 of the 11 loci). In addition,

25 of the 41 males were pupae, indicating that they had been produced within their parental colony. The most likely mechanism for this mode of clonal reproduction is the paternal elimination of the maternal genome in the egg. Accordingly, the resulting haploid males produced would lack maternal genes and would have a genotype identical to the sperm stored in the queens' spermathecae. The alternative mechanism of androgenesis (the fusion of two sperm nuclei, leading to all-paternal diploid males) can be ruled out because flow cytometry analysis conducted on 10 male heads revealed that they were all haploid (S. Aron, personal communication).

By using alternative modes of reproduction for the queen and worker castes, queens can increase the transmission rate of their genes to their reproductive female offspring while maintaining genetic diversity in the worker force. The fact that sexual reproduction has been retained to produce workers indicates that sexual reproduction might have important benefits for colony function, for example through increased defence against parasites, more efficient division of labour and an increased range of environmental conditions that a colony can tolerate⁷⁻⁹. Thus, if queens were also to produce workers clonally, all females within colonies would have identical genomes, whereas the mode of reproduction that we have uncovered effectively leads to levels of colony genetic diversity identical to that expected in a typical colony headed by one queen mated with one male. A similar situation of conditional use of sex by queens has been reported in another ant, *Cataglyphis cursor*¹⁰. Interestingly, in *C. cursor* and *W. auropunctata* the cost of asexual

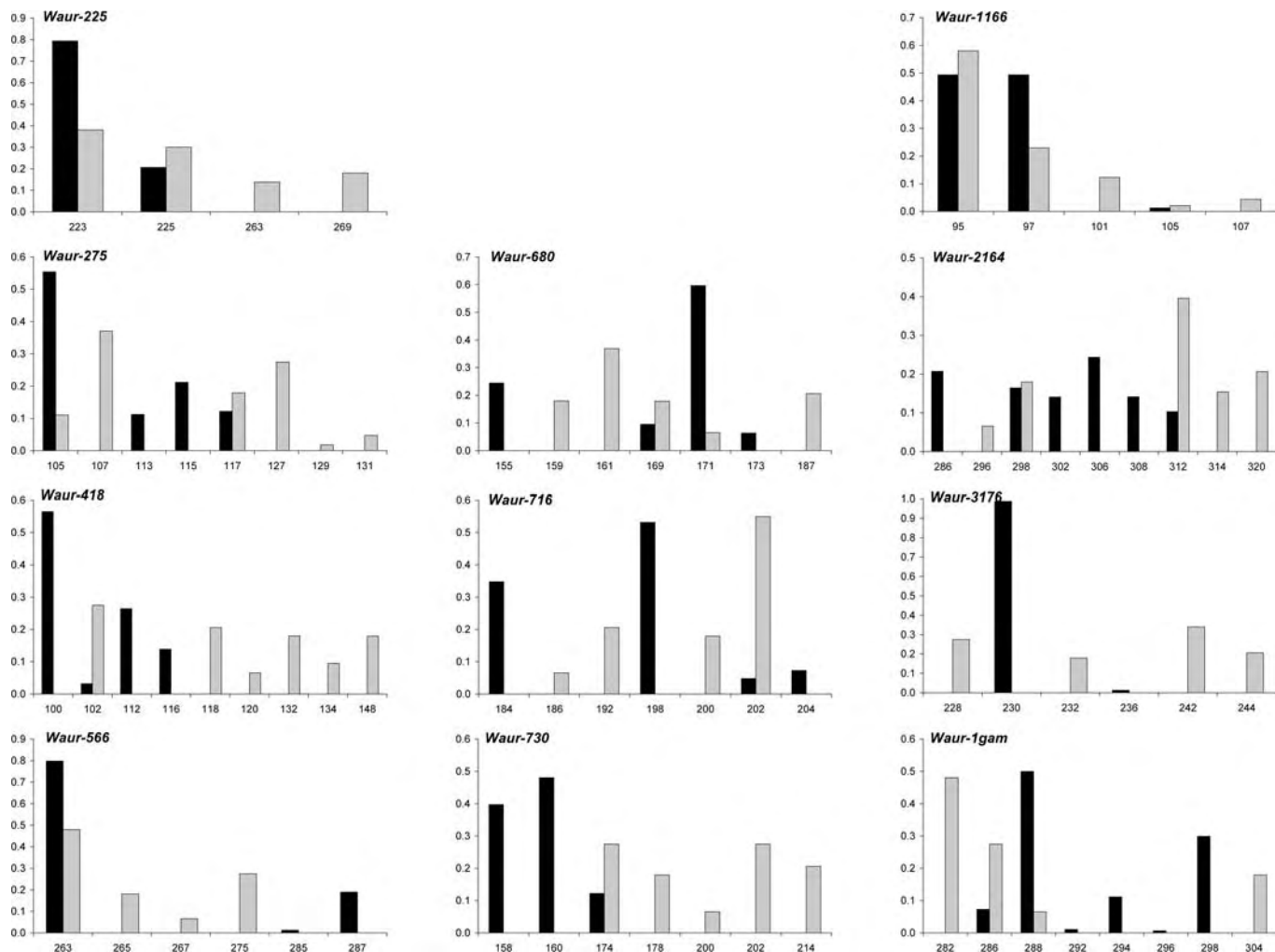


Figure 3 | Allele frequencies for queens (black bars) and their mates (grey bars) at each of the 11 microsatellite loci.

reproduction might be lower than in most other ants because in these two species queens do not go through a stage of independent colony founding in which queens initiate a new colony without the help and protection of workers¹⁰. However, an important difference is that parent–offspring analyses revealed numerous cases of crossing-over events in *C. cursor* as a result of automictic parthenogenesis with central fusion, a process in which two of the four products of meiosis merge. In contrast, parthenogenesis seems to be only, or at least mostly, ameiotic in *W. auropunctata* as indicated by the fact that the vast majority of queens from the same nest and sometimes all the queens from the same site have the same genotype. The relatively high level of queen heterozygosity (mean $\pm$ s.e.m. 0.619 ± 0.028) also shows a very low rate or the complete absence of recombination, because crossing-over should result in offspring becoming homozygous for one of the two maternal alleles and a rapid decrease in heterozygosity¹¹.

To our knowledge there has been only one other report of clonal reproduction by males in the animal kingdom, in which it was found that 3 of 61 queens homozygous for a recessive (cordovan) mutation inducing a brown instead of the wild-type black colour produced some brown males¹². This showed convincingly that some males could not originate from unfertilized eggs produced by queens, but it is possible that these males were diploid, as sometimes occurs in Hymenoptera¹³. Some support for the occurrence of diploid males comes from the observation that the three queens also produced some gynandromorphic males with parts of cordovan (brown) and dark coloured tissue, a pattern best explained by diploid males experiencing partial reduction of ploidy level during development. Although the available data do not allow us to conclude whether clonal reproduction by males does really occur in honeybees, it is important to note that this mode of reproduction might have remained unnoticed in social Hymenoptera because of the paucity of suitable parent–offspring genetic analyses. Good candidates for male clonal reproduction are species in which the reproduction of workers has been inferred on the basis that males harboured alleles absent in queens but present in their mates (and thus in workers).

The elimination of one of the two parental genomes during meiosis has been described in fishes, amphibians and several insect species^{14,15}. However, these cases invariably involve the elimination of the paternal genome. A case of male induction by genome elimination has also been reported in a haplodiploid hymenopteran, the wasp *Nasonia vitripennis*, in which a paternal-sex-ratio (*psr*) chromosome induces the supercondensation and destruction of the paternal chromosomes (except *psr*) in early fertilized eggs^{16,17}. The selfish effect of this B chromosome is to convert diploid eggs, which would have normally developed into females, into haploid males. However, in this system it is also the paternal and not the maternal genome that is eliminated. Thus, the *W. auropunctata* system described here is unique in that clonal reproduction occurs by the transmission of the paternal and not the maternal genome. This system might have evolved in response to the obligate clonal production of queens as a mechanism for males to achieve some reproductive success.

The occurrence of clonal reproduction by both males and queens has important consequences for the apportionment of genetic variability and genome evolution. Because genes are transmitted only between individuals of the same sex, there is effectively no gene flow between the male and queen gene pools. Genetic differentiation between the male and queen genomes can thus persist and accumulate, as demonstrated by the fact that the genomic compositions of queens differed notably from those of males (Fig. 3). Allele frequencies were significantly different between sexes (Fisher's exact tests, $P < 0.00001$ for each of the 11 loci), with some loci such as *Waur-3176* being diagnostic. A dendrogram analysis also revealed that males and queens clustered on separate branches of the tree (Fig. 2). There was a complete segregation of the queens' and the

males' genomes. Accordingly, the heterozygosity of the sexually produced workers (mean $\pm$ s.e.m. 0.858 ± 0.013 ; range 0.693–0.989) was significantly higher than that of queens (permutation test, $P = 10^{-5}$), and their genotypes showed a significant departure from Hardy–Weinberg equilibrium, with a strong excess of heterozygous genotypes at all loci ($P < 10^{-5}$ at each locus).

This study shows that, in the evolutionary battle of opposing sexes, *W. auropunctata* has evolved an unusual mode of reproduction with queens circumventing the twofold cost of sexual reproduction by transferring all their genes to the reproductive females while males thwart queens by also clonally transmitting their genomes to sons. Although the male and female genomes come together in workers, this does not translate into any genetic exchange because workers are completely sterile¹⁸. As a result, the male and female genomes are completely segregated and form two distinct genetic lineages. These findings show that haplodiploidy and the caste-determination system provide grounds for the evolution of extraordinary genetic systems and that sexual conflicts are central in evolution with the potential for shaping various interactions between the sexes and their gametes^{19–23}.

METHODS

Sampling. Queens, gynes (winged queens), males and workers of *W. auropunctata* were sampled in March 2004 within their native range, in French Guiana. Thirty-four nests were surveyed and collected at five sites (Fig. 1): two coffee-tree plantations (A, $05^{\circ}17'20.82''$ N, $52^{\circ}55'11.40''$ W; B, $05^{\circ}17'16.02''$ N, $52^{\circ}55'04.08''$ W), a sand-pit (C, $05^{\circ}16'13.92''$ N, $52^{\circ}55'02.70''$ W), an old encampment (D, $05^{\circ}04'21.18''$ N, $52^{\circ}01'47.16''$ W) and a quarry (E, $05^{\circ}04'16.81''$ N, $52^{\circ}02'44.70''$ W). In addition, a sample of queens and males collected in New Caledonia along a forest road ($20^{\circ}33'16.71''$ S, $164^{\circ}47'53.23''$ E), in a coffee-tree plantation ($22^{\circ}01'57.05''$ S, $166^{\circ}16'12.09''$ E) and in a rainforest ($22^{\circ}10'18.13''$ S, $166^{\circ}45'37.16''$ E) were included in this study.

Nuclear genotyping. To isolate sperm DNA, the queen's abdomen was dissected as described in ref. 24. Whole individual ants were ground in cetyltrimethylammonium bromide (CTAB) solution. DNA was extracted and purified in accordance with standard CTAB-based protocols. Queens, males, gynes, workers and seminal fluid were genotyped at 11 highly polymorphic microsatellite loci (*Waur-225*, *Waur-275*, *Waur-418*, *Waur-566*, *Waur-680*, *Waur-716*, *Waur-730*, *Waur-1166*, *Waur-2164*, *Waur-3176* and *Waur-1gam*; see detailed protocol in ref. 25.).

Microsatellite analyses. Tests for departure from Hardy–Weinberg equilibrium and genic differentiation between male and female gene pools were assessed with the exact probability tests implemented in GENEPOP 3.2a (ref. 26). Pairwise individual allele-shared distances were estimated over all loci as

$$ASD_{XY} = \frac{1}{L} \sum_{l=1}^L 1 - [(n_{lX \rightarrow Y} + n_{lY \rightarrow X}) / (G_X + G_Y)]$$

where L is the number of genotyped loci, G_X and G_Y are the numbers of gene copies in individuals X and Y ($G = 2$ for diploid genomes and $G = 1$ for haploid genomes), respectively, and $n_{lX \rightarrow Y}$ ($n_{lY \rightarrow X}$) is the number of gene copies at locus l in individual X (Y) for which the allelic state is also observed (that is, shared) in individual Y (X). Calculations of allele-shared distances between pairs of individuals and construction of neighbour-joining dendrograms were performed with the program TreeMaker (S. Piry, personal communication). Dendrograms were constructed using the TreeView 1.6.6 program²⁷.

Received 5 February; accepted 4 May 2005.

- Hamilton, W. D. Extraordinary sex ratios. *Science* **156**, 477–488 (1967).
- Eberhard, W. G. Evolutionary consequences of intracellular organelle competition. *Q. Rev. Biol.* **55**, 231–249 (1980).
- Hurst, L. D. Intragenomic conflict as an evolutionary force. *Proc. R. Soc. Lond. B* **248**, 135–140 (1992).
- Pomiankowski, A. in *Levels of Selection in Evolution* (ed. Keller, L.) 121–152 (Princeton Univ. Press, Princeton, 1999).
- Hölldobler, B. & Wilson, E. O. *The Ants* (Springer, Berlin, 1990).
- Torres, J. A., Snelling, R. R. & Canals, M. Seasonal and nocturnal periodicities in ant nuptial flights in the Tropics (Hymenoptera: Formicidae). *Sociobiology* **37**, 601–626 (2001).
- Sherman, P. W., Seeley, T. D. & Reeve, H. K. Parasites, pathogens, and polyandry in social Hymenoptera. *Am. Nat.* **131**, 602–610 (1988).

8. Crozier, R. H. & Page, R. E. On being the right size: male contributions and multiple mating in social Hymenoptera. *Behav. Ecol. Sociobiol.* **18**, 105–115 (1985).
9. Cole, B. J. & Wiernasz, D. C. The selective advantage of low relatedness. *Science* **285**, 891–893 (1999).
10. Pearcy, M., Aron, S., Doums, C. & Keller, L. Conditional use of sex and parthenogenesis for worker and queen production in ants. *Science* **306**, 1780–1783 (2004).
11. Baudry, E. *et al.* Whole-genome scan in thelytokous-laying workers of the Cape honeybee (*Apis mellifera capensis*): central fusion, reduced recombination rates and centromere mapping using half-tetrad analysis. *Genetics* **167**, 243–252 (2004).
12. Koeniger, N., Hemmeling, C. & Yoshida, T. Drones as sons of drones in *Apis mellifera*. *Apidologie (Celle)* **20**, 391–394 (1989).
13. Cook, J. M. & Crozier, R. H. Sex determination and population biology in the Hymenoptera. *Trends Ecol. Evol.* **10**, 281–286 (1995).
14. Simon, J.-C., Delmotte, F., Risse, C. & Crease, T. Phylogenetic relationships between parthenogens and their sexual relatives: the possible routes to parthenogenesis in animals. *Biol. J. Linn. Soc.* **79**, 151–163 (2003).
15. Normark, B. B. The evolution of alternative genetic systems in insects. *Annu. Rev. Entomol.* **48**, 397–423 (2003).
16. Werren, J. H., Nur, U. & Eickbush, D. An extrachromosomal factor causing loss of paternal chromosomes. *Nature* **327**, 75–76 (1987).
17. Nur, U., Werren, J. H., Eickbush, D. G., Burke, W. D. & Eickbush, T. H. A 'selfish' B chromosome that enhances its transmission by eliminating the paternal genome. *Science* **240**, 512–514 (1988).
18. Ulloa-Chacon, P. & Cherix, D. Quelques aspects de la biologie de *Wasmannia auropunctata* (Roger) (Hymenoptera, Formicidae). *Actes Colloq. Insectes Sociaux* **4**, 177–184 (1988).
19. Rice, W. R. Sexually antagonistic male adaptation triggered by experimental arrest of female evolution. *Nature* **381**, 232–234 (1996).
20. Swanson, W. J., Yang, Z., Wolfner, M. F. & Aquadro, C. F. Positive Darwinian selection drives the evolution of several female reproductive proteins in mammals. *Proc. Natl Acad. Sci. USA* **98**, 2509–2514 (2001).
21. Arnqvist, G. & Rowe, L. Antagonistic coevolution between the sexes in a group of insects. *Nature* **415**, 787–789 (2002).
22. Helms-Cahan, S. & Keller, L. Complex hybrid origin of genetic caste determination in harvester ants. *Nature* **424**, 306–309 (2003).
23. Parker, J. D. A major evolutionary transition to more than two sexes? *Trends Ecol. Evol.* **19**, 83–86 (2004).
24. Chapuisat, M. Mating frequency of ant queens with alternative dispersal strategies, as revealed by microsatellite analysis of sperm. *Mol. Ecol.* **7**, 1097–1105 (1998).
25. Fournier, D. *et al.* Characterization and PCR multiplexing of polymorphic microsatellite loci for the invasive ant *Wasmannia auropunctata*. *Mol. Ecol. Notes* **5**, 239–242 (2005).
26. Raymond, M. & Rousset, F. GENEPOP (version 1.2): population genetics software for exact tests and ecumenicism. *J. Hered.* **86**, 248–249 (1995).
27. Page, R. D. M. TreeView: an application to display phylogenetic trees on personal computers. *Comput. Appl. Biosci.* **12**, 357–358 (1996).

Acknowledgements We are grateful to the Laboratoire Environnement de Petit Saut (EDF-CNEH) for hosting us during field work, the Centre de Coopération Internationale en Recherche Agronomique pour le Développement (CIRAD) of French Guiana for allowing access to plantations, S. Aron for performing flow cytometry analysis, and M. Chapuisat, P. Christie, D. Haig, R. Hammond, S. Helms Cahan, K. Parker and D. Queller for comments on the manuscript. The work was supported by grants from the French Ministère de l'Écologie et du Développement Durable to A.E. and the Swiss NSF to L.K.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to D.F. (denis.fournier@ulb.ac.be) or A.E. (estoup@ensam.inra.fr).

Orofacial somatomotor responses in the macaque monkey homologue of Broca's area

Michael Petrides¹, Geneviève Cadoret¹ & Scott Mackey¹

In the ventrolateral frontal lobe of the human brain there is a distinct entity, cytoarchitectonic area 44 (Broca's area), which is crucial in speech production^{1–4}. There has been controversy^{5,6} over whether monkeys possess an area comparable to human area 44. We have addressed this question in the macaque monkey by combining quantitative architectonic analysis of the cortical areas within the ventrolateral frontal region with electrophysiological recording of neuron activity and electrical intracortical microstimulation. Here we show that, immediately in front of the ventral part of the agranular premotor cortical area 6, there is a distinct cortical area that is architectonically comparable to human area 44 and that this monkey area 44 is involved with the orofacial musculature. We suggest that area 44 might have evolved originally as an area exercising high-level control over orofacial actions, including those related to communicative acts, and that, in the human brain, area 44 eventually also came to control certain aspects of the speech act.

In the human brain, in the ventrolateral frontal cortex, there is a distinct region that exercises critical control over the motor aspects of speech production. This region is often referred to as Broca's area^{1–4}. Some investigators⁷ have used the term Broca's area to refer to architectonic areas 44 and 45, whereas others⁸ have restricted the term to area 44. Although several attempts have been made to specify more precisely the critical zone for speech within this general region on the basis of clinical–anatomical correlation studies^{1,8,9}, such endeavours have been of limited success because of the difficulty in obtaining lesions confined to a specific area. The best evidence has been obtained from electrical stimulation of the cerebral cortex under local anaesthesia during brain surgery to establish, in a particular patient, the precise part of the left hemisphere that is critical for speech. The critical region is considered to be the part of the cortex from which speech production can be interrupted by the application of electrical stimulation. Speech arrest occurs most reliably from stimulation of the pars opercularis^{1–4}, which is occupied by area 44 (refs 7, 10). Area 44 is bounded caudally by the ventral part of premotor area 6, which controls the orofacial musculature, and rostrally by area 45 (Fig. 1a, b).

Although it is not disputed that, in the macaque monkey as in the human, the most ventral part of the precentral gyrus (areas 4 and 6) is involved with the motor control of the orofacial musculature, there is considerable confusion and debate as to whether, just rostral to the premotor area 6, there might be a cortical region corresponding to area 44 of the human brain^{5,6}. This debate has become of particular concern in recent years with the discovery of a class of neurons known as 'mirror neurons' in part of the ventral premotor cortical area 6 (also known as area F5c) of the macaque monkey^{5,11,12}. These neurons become active both when the monkey performs a particular action and when the monkey observes a similar action being performed by another individual. Because there has been considerable theoretical interest in the possibility that the mirror-neuron

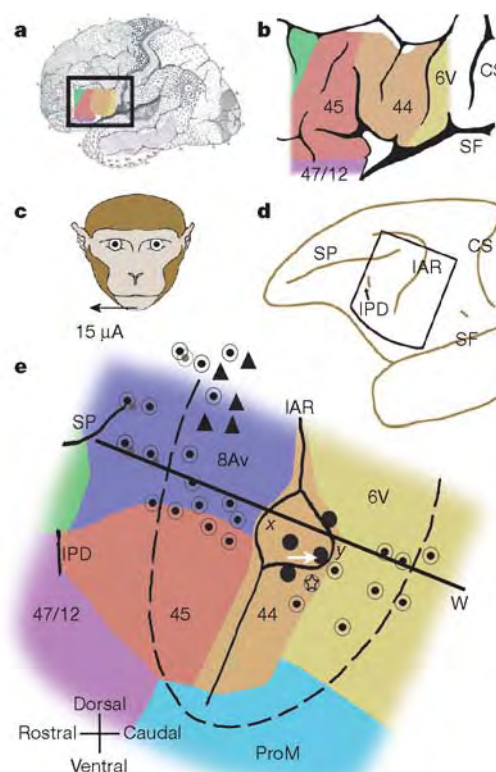


Figure 1 | Region of interest and location of experimental results in the monkey. **a**, Architectonic map of the human cerebral cortex by Brodmann¹⁰ with the areas in the caudal part of the ventrolateral frontal lobe coloured and outlined. **b**, Close-up of outlined area in **a**. **c**, Example of orofacial response: intracortical microstimulation at the site indicated by the white arrow in **e** evoked jaw muscle contraction resulting in a lateral motion of the jaw. **d**, The region of interest as traced from a photograph of the left hemisphere of monkey 1. **e**, Diagrammatic representation of the unfolded inferior arcuate sulcus showing the location of architectonic areas 6V, 44, 45, 8Av, 47/12 and ProM. The reconstruction was generated by measuring, on the digital images, the distance between the fundus and the rostral and caudal lips of the sulcus along layer IV of the cortex. The lips of the sulcus are represented by the dashed line, and the fundus of the sulcus by the solid curved line. The floor of the sulcus flattens at a certain level. Points *x* and *y* on line W indicate this flattening and identify the same locations in Fig. 2a. The symbols indicate the recording sites and the evoked responses: triangles, oculomotor response; dots in circles, no response; filled circles, orofacial response; star in circle, orofacial plus hand movement response. CS, central sulcus; IAR, inferior arcuate sulcus; IPD, infraprecipital dimple; SF, sylvian fissure; SP, sulcus principalis.

¹Montreal Neurological Institute, McGill University, 3801 University Street, Montreal, Quebec H3A 2B4, Canada.

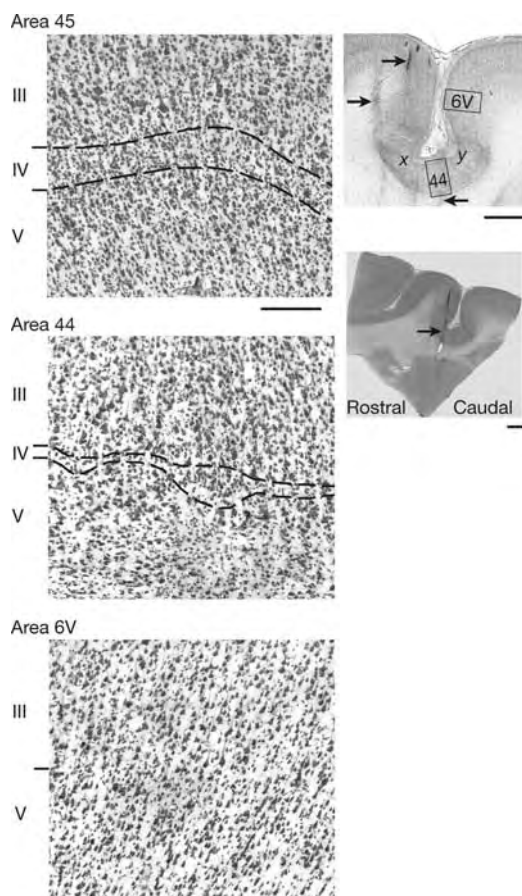


Figure 2 | Photomicrographs of areas 45, 44 and 6V focused on granular layer IV and adjacent layers III and V. Granular layer IV is marked by the black dashed lines. The photomicrographs of areas 6V and 44 were taken from the section shown at top right (represented by line W in Fig. 1e). The photomicrograph of area 45 was acquired from a more ventral section. In the sections on the right, note the electrode tracks (arrowed) that reached the fundus of the sulcus. Scale bars, 200 μ m (left); 1 mm (right).

system might be important for the evolution of language⁵, it has been suggested that the homologue of Broca's area in the monkey might be the ventral premotor cortical area F5 (part of area 6) within which the mirror neurons were discovered^{5,6}. However, this area is agranular, whereas area 44 (Broca's area) in the human brain is a dysgranular architectonic entity that lies rostral to the ventral agranular premotor cortex^{7,10}. The key question is therefore: Is there a cortical area immediately in front of premotor cortical area 6 of the macaque monkey brain that is comparable to area 44 of the human brain? Such an area should have the following three properties: first, it should exhibit the key architectonic characteristics of area 44 of the human brain; second, it should be bounded topographically by the same architectonic areas as in the human brain; and third, it should be involved, at least, with the orofacial musculature. We addressed the above question in the macaque monkey by combining quantitative architectonic analysis of the cortical areas within the inferior arcuate sulcus, which defines the transition between the ventral precentral motor region and the ventral prefrontal cortex (Fig. 1d), with electrophysiological recording of neuron activity and electrical intracortical microstimulation.

There are two essential cytoarchitectonic characteristics that discriminate between areas 45, 44 and 6V in the ventrolateral frontal cortex of the human brain⁷. The first is that area 45 has clusters of unusually large neurons in the deeper part of layer III that are even larger than the largest neurons of layer V; the second is that the granular layer IV is well developed in granular area 45, it is

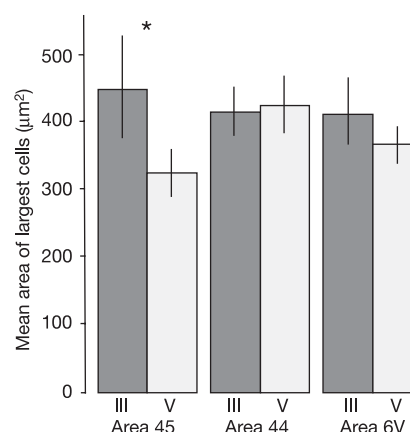


Figure 3 | Mean area of the largest cells in layers III and V of areas 45, 44 and 6V. Areas were measured on the digital images of six histological sections from both monkeys. The error bars indicate s.d. The asterisk indicates that the asymmetry in the size of the largest neurons in layers III and V in area 45 was statistically significant. This asymmetry, combined with a well-developed granular layer IV (see Fig. 4), is the typical characteristic of area 45 in both the human and the monkey brain.

rudimentary and irregular in dysgranular area 44, and is absent in agranular area 6V. The present architectonic analysis in the monkey ventrolateral frontal region showed that, as in the human brain (Fig. 1a, b), the ventral part of agranular area 6 (Fig. 2) is replaced, rostrally, by dysgranular area 44, which exhibits a rudimentary and irregular layer IV (Fig. 2). Area 44, which lies on the deeper part of the caudal bank and the fundus of the inferior arcuate sulcus, is replaced, in turn, by granular area 45 (Fig. 2) in the rostral bank of this sulcus (Fig. 1e). Figure 3 shows the statistically significant asymmetry between the largest neurons in inner layer III and the largest neurons in layer V in area 45 (see Methods). Figure 4 shows the statistically significant change in granularity in layer IV across the three ventrolateral areas 6, 44 and 45 (see Methods). In monkeys 1 and 2, the maximum rostrocaudal width of area 44, measured along layer IV, was 4 mm and 9 mm, respectively. Thus, in both the human and the monkey ventrolateral frontal cortex, proceeding in a caudal to rostral direction, agranular premotor area 6 is replaced by dysgranular area 44, which, in turn, is replaced by granular area 45 (Fig. 1a, b, e). The dysgranular area 44 has also been shown to lie rostral to the agranular ventral premotor area 6 in our closest living primate relative, the chimpanzee¹³. However, this study was based on qualitative rather than quantitative criteria, and no electrophysiological work was performed to confirm the involvement of area 44 with the orofacial musculature. An earlier study had suggested that there is an asymmetry in area 44 in the brain of the chimpanzee and gorilla from measurements of gross sulcal morphology visible on magnetic resonance images¹⁴. Nevertheless, this conclusion must be treated with caution because area 44 was not identified on the basis of architectonic analysis¹⁴.

In the electrophysiological part of the present investigation, the region of the inferior arcuate sulcus was explored in 52 recording and stimulation sites (39 sites in monkey 1, shown in Fig. 1e, and 13 in monkey 2). At these recording sites, neuron receptive field and behavioural response to intracortical microstimulation were tested (see Methods). Saccadic eye movements were evoked in response to microstimulation at seven sites at thresholds varying from 45 to 80 μ A. Saccades were in the 180° range from upwards to downwards in the contralateral direction. As expected^{15–17}, intracortical microstimulation evoked saccadic eye movements close to the dorsal portion of the rostral bank of the inferior arcuate sulcus, at the level of the sulcus principalis, namely within the frontal eye field (Fig. 1e). Architectonic examination showed that these oculomotor responses were located in area 8Av. No oculomotor responses were

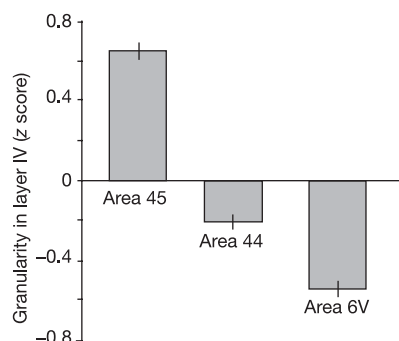


Figure 4 | Quantitative assessment of the relative granularity of layer IV in architectonic areas 45, 44 and 6V. The average z score of the measure of granularity of layer IV in areas 45, 44 and 6V from monkey 1 is shown (see Methods). The error bars indicate s.e.m. The granularity is highest in granular area 45, lowest in agranular area 6V and at an intermediate level in dysgranular area 44. These differences were statistically significant (see Methods).

observed from stimulation in the ventral part of the rostral bank of the inferior arcuate sulcus where the cortex exhibits the characteristics of area 45 of the human brain, namely the presence of clusters of large deeply stained pyramidal neurons in the deeper part of layer III combined with a well-developed layer IV and neurons of medium to small size in layer V (Figs 2 and 3) (see Supplementary Discussion).

Orofacial responses were observed at eight sites (five in monkey 1 and three in monkey 2). These responses were located at a depth of 4–6 mm from the surface of the brain. All the sites from which orofacial responses were obtained were located within architectonic area 44, which occupies the ventral part of the inferior arcuate sulcus within the fundus and the deeper half of the caudal bank of the sulcus (Fig. 1e). One site in area 44 showed combined orofacial and hand movement responses. At seven of the eight sites with orofacial responses in area 44, motor orofacial responses were evoked by intracortical microstimulation at thresholds varying from 15 to 80 μ A. These responses were essentially jaw movements, such as jaw closing, sequences of jaw opening and closing, and jaw movements with horizontal vector components. In one case, microstimulation evoked bilateral upper lip movements. In two cases, jaw motor responses were associated with orofacial mechanoreceptive fields. One neuron responded to firm pressure over the belly of the masseter muscle, and microstimulation (60 μ A) evoked a jaw closing. Another neuron responded to light tactile stimulation of the intraoral soft tissue and the tongue dorsum, and microstimulation (15 μ A) evoked jaw muscle contractions resulting in lateral motion of the jaw (Fig. 1c). In one site, electrical stimulation evoked no response, but the recorded neuronal activity was clearly modulated by passive retrusion of the tongue. In one monkey, microstimulation in area 44 repeatedly evoked respiratory responses, such as brief forced expirations. We should emphasize here that the aim of the present electrophysiological study was limited, namely to establish whether neurons in architectonically defined area 44 are sensitive to orofacial stimuli and whether their electrical stimulation can evoke orofacial movements that provide further support to the architectonic conclusions. No effort was made to examine in detail the properties of neurons in this area and compare them with those of area 6V.

The present study has established that a cortical area comparable in architecture to human area 44 exists in the macaque monkey immediately in front of premotor cortical area 6V and that it is involved with the orofacial musculature. The lack of an outgroup comparison limits our ability to provide further inferential evidence that area 44 in the human and the macaque monkey brain reflect shared common ancestry. Area 44 in the monkey lies rostral to the convexity of the premotor cortex where the mirror neurons were recorded (area F5c)⁵. Thus, the involvement (if any) of area 44 in the

mirror neuron system remains to be established. Furthermore, it has recently been argued that mirror neurons cannot provide a basis for an essential structural relation in human language, namely the bi-directional arbitrary mapping between sound and meanings¹⁸. In the human brain, area 44 is involved with the motor aspects of speech production^{1–4}. Studies of the effects of lesions that are more or less restricted to area 44 have yielded an apraxia of speech (that is, a problem with the motor aspects of speech production), and not the classic full-blown aphasic syndrome including a major disruption of syntax, which was previously thought to be the result of damage to Broca's area^{8,9}.

The position of area 44 in the monkey and the human brain immediately in front of the premotor cortex and its orofacial responses to microstimulation suggest that it is an essential link in a series of cortical areas^{19,20} that control the orofacial musculature. Area 44 is in a unique position to communicate directly both with the agranular premotor cortex (area 6V) that is controlling the orofacial musculature and with area 45, which has been shown to be involved in the active retrieval of information from memory in both the monkey and the human brain^{21–23} (including verbal episodic information²⁴ and semantic information²⁵ in the left hemisphere of the human brain). We suggest that area 44 might have evolved originally as an area exercising high-level control over orofacial and other actions, including those related to communicative acts, and that, in the human brain, with the evolution of language, area 44 in the left hemisphere eventually also came to control the motor aspects of the speech act.

METHODS

Animal preparation and data collection. The electrophysiological mapping study was conducted on two adult monkeys (*Macaca fascicularis*) weighing 5.0 and 6.5 kg, respectively. A stainless steel recording chamber was implanted, under general anaesthesia, over the inferior arcuate sulcus. During recording of neuronal activity and intracortical microstimulation, the monkeys were anaesthetized with Telazol (tiletamine HCl and zolazepam HCl, 13 mg kg⁻¹), a dissociative agent with no significant effect on electrical stimulation thresholds for limb and eye movements^{26–28} or on the velocity and duration of evoked eye movements²⁹. Twice a week, glass-insulated tungsten microelectrodes (impedance 0.5–1.7 M Ω , measured at 1 kHz frequency), mounted on a hydraulic microdrive, were advanced into the region of interest until neuronal activity was detected. After amplification, neuronal activity was monitored on an oscilloscope and individual action potentials were isolated with a voltage discriminator. Isolated neurons were examined to identify their receptive fields. Receptive field testing consisted of three processes: first, visual stimulation by moving a target in different directions in front of the monkey's eyes; second, light tactile stimulation by applying a cotton applicator stick to different parts of the face (skin, lips, tongue, internal wall of the mouth) and to the glabrous and hairy parts of the hands; and third, manual pressure or stretch application on the orofacial muscles and the hand muscles. A neuron was considered to be responsive to any one of these stimuli if there was a consistent increase or decrease in neuronal activity evoked by repetitive presentation (5–10 times) of the stimulus. After sensory testing, microstimulation was applied without moving the position of the microelectrode. Microstimulation consisted of unipolar cathodal pulses, delivered at a frequency of 300 Hz with train durations ranging from 100 to 500 ms. Current levels from 10 to 100 μ A were used to search for evoked responses in the inferior arcuate sulcus. The minimum intensity evoking eye movements or orofacial motor responses was considered to be the stimulation threshold. Behavioural responses were recorded by two independent observers. Before the animals were killed, several electrolytic lesions (30 μ A cathodal current for 30 s) were made to help in reconstructing the penetration sites and to correlate the electrophysiological findings with the quantitative architectonic analysis.

Quantitative architectonic analysis. To examine quantitatively the asymmetry between the largest neurons in layers III and V in area 45 versus the lack of asymmetry in areas 44 and 6V, a standard sampling window (770 μ m $\times$ 930 μ m) was superimposed at three locations in the three architectonic areas on each of six images (that is, $3 \times 3 \times 3 \times 2 = 54$ windows) acquired from a series of histological sections (see Supplementary Methods). The windows were centred on layer IV and included the inner part of layer III and the outer part of layer V, namely the parts of layers III and V that contain their largest neurons. In area 6,

which lacks a definite layer IV, the sampling windows were oriented on the boundary between layers III and V. All large neurons within each window were measured by manually tracing their perimeter and extracting the enclosed area with Northern Eclipse image analysis software (Empix Imaging). The ten neurons with the greatest area from both layers III and V in every sampling window were entered as values in a two-way analysis of variance. There was a significant layer $\times$ area interaction (monkey 1: $F = 23.03$, d.f. = 2, 534, $P < 0.05$; monkey 2: $F = 47.99$, d.f. = 2, 534, $P < 0.05$). Tests of simple main effects showed that, in area 45, the difference in size of cells between layers III and V was statistically significant (monkey 1: $F = 39.3$, d.f. = 1, 534, $P < 0.01$; monkey 2: $F = 115.83$, d.f. = 1, 534, $P < 0.01$).

To assess quantitatively the differences in granularity between architectonic areas 44, 45 and 6, a region of interest including all of layer IV (the granular layer) was selected on digital photomicrographs (see Supplementary Methods). Cells within the region of interest were segmented from the cortical matrix by thresholding and then classified on the basis of their area. The locations of individual cells within the size range of typical granule neurons were coded by a small circle on a separate digital layer. To facilitate sampling, the layer coding the location of granule sized cells was convolved with a gaussian blurring kernel and then sampled by a series of equidistant traverses arranged perpendicular to the cortical layers³⁰ (see Supplementary Methods). The concentration of granule-sized cells in layer IV was recorded from each profile. These values were converted to z scores on a section by section basis, grouped according to architectonic area, and compared statistically by a one-way analysis of variance in each monkey. The difference in granularity between architectonic areas was determined as being significant in both monkeys ($F = 67.37$, d.f. = 2, 370, $P < 0.001$; $F = 487.93$, d.f. = 2, 569, $P < 0.001$). Post-hoc comparisons with Tukey's Honestly Significant Difference test indicated that all areas differed significantly from each other ($P < 0.01$).

Received 8 March; accepted 12 April 2005.

- Penfield, W. & Roberts, L. *Speech and Brain Mechanisms* (Princeton Univ. Press, Princeton, New Jersey, 1959).
- Rasmussen, T. & Milner, B. in *Cerebral Localization* (eds Zulch, K. J., Creutzfeldt, O. & Galbraith, G. C.) 238–257 (Springer, Berlin, 1975).
- Ojemann, G. A., Ojemann, J. G., Lettich, E. & Berger, M. S. Cortical language localization in left, dominant hemisphere. An electrical stimulation mapping investigation in 117 patients. *J. Neurosurg.* **71**, 316–326 (1989).
- Duffau, H. et al. The role of dominant premotor cortex in language: a study using intraoperative functional mapping in awake patients. *Neuroimage* **20**, 1903–1914 (2003).
- Rizzolatti, G. & Craighero, L. The mirror-neuron system. *Annu. Rev. Neurosci.* **27**, 169–192 (2004).
- Matelli, M., Luppino, G., Geyer, S. & Zilles, K. in *The Human Nervous System* (eds Paxinos, G. & Mai, J. K.) 973–996 (Elsevier Academic, San Diego, 2004).
- Amunts, K. et al. Broca's region revisited: Cytoarchitecture and intersubject variability. *J. Comp. Neurol.* **412**, 319–341 (1999).
- Mohr, J. P. in *Studies in Neurolinguistics* (eds Winitaker, H. & Winitaker, N. A.) 201–235 (Academic, New York, 1976).
- Poeppel, D. & Hickok, G. Towards a new functional anatomy of language. *Cognition* **92**, 1–12 (2004).
- Brodman, K. *Vergleichende Lokalisationslehre der Grosshirnrinde in ihren Prinzipien dargestellt auf Grund des Zellenbaues* (Barth, Leipzig, 1909).
- Rizzolatti, G., Fadiga, L. & Gallese, V. Premotor cortex and the recognition of motor actions. *Cogn. Brain Res.* **3**, 131–141 (1996).
- Ferrari, P. F., Gallese, V., Rizzolatti, G. & Fogassi, L. Mirror neurons responding to the observation of ingestive and communicative mouth actions in the monkey ventral premotor cortex. *Eur. J. Neurosci.* **17**, 1703–1714 (2003).
- Sherwood, C. C., Broadfield, D. C., Holloway, R. L., Gannon, P. J. & Hof, P. R. Variability in Broca's area homologue in African Great Apes: Implications for language evolution. *Anat. Rec. A* **271**, 276–285 (2003).
- Cantalupo, C. & Hopkins, W. D. Asymmetric Broca's area in great apes. *Nature* **414**, 505 (2001).
- Bruce, C. J., Goldberg, M. E., Stanton, G. B. & Bushnell, M. C. Primate frontal eye fields. II. Physiological and anatomical correlates of electrically evoked eye movements. *J. Neurophysiol.* **54**, 714–734 (1985).
- Huerta, M. F., Krubitzer, L. A. & Kaas, J. H. Frontal eye fields as defined by intracortical microstimulation in squirrel monkeys, owl monkeys, and macaque monkeys. II. Cortical connections. *J. Comp. Neurol.* **271**, 473–492 (1987).
- Stanton, G. B., Deng, S.-Y., Goldberg, M. E. & McMullen, N. T. Cytoarchitectural characteristic of the frontal eye fields in macaque monkeys. *J. Comp. Neurol.* **282**, 415–427 (1989).
- Hurford, J. in *Evolution of Communication Systems: A Comparative Approach* (eds Kimbrough Oller, D., Griebel, M. & Plunkett, K.) 297–313 (MIT Press, Cambridge, Massachusetts, 2004).
- Matelli, M., Camarda, R., Glickstein, M. & Rizzolatti, G. Afferent and efferent projections of the inferior area 6 in the macaque monkey. *J. Comp. Neurol.* **251**, 281–298 (1986).
- Tokuno, H., Takada, M., Nambu, A. & Inase, M. Reevaluation of ipsilateral corticocortical inputs to the orofacial region of the primary motor cortex in the macaque monkey. *J. Comp. Neurol.* **389**, 34–48 (1997).
- Petrides, M. The mid-ventrolateral prefrontal cortex and active mnemonic retrieval. *Neurobiol. Learn. Mem.* **78**, 528–538 (2002).
- Cadoret, G. & Petrides, M. Neuronal mechanisms underlying active retrieval processing in the monkey (Abstract Viewer/Itinerary Planner. Soc. Neurosci. Prog. No. 324, 3, 2004).
- Cadoret, G., Pike, G. B. & Petrides, M. Selective activation of the ventrolateral prefrontal cortex in the human brain during active retrieval processing. *Eur. J. Neurosci.* **14**, 1164–1170 (2001).
- Petrides, M., Alivisatos, B., Meyer, E. & Evans, A. C. Functional activation of the human ventrolateral frontal cortex during mnemonic retrieval of verbal information. *Proc. Natl Acad. Sci. USA* **92**, 5803–5807 (1993).
- Amunts, K. et al. Analysis of neural mechanisms underlying verbal fluency in cytoarchitecturally defined stereotaxic space—the roles of Brodmann areas 44 and 45. *Neuroimage* **22**, 42–56 (2004).
- Hoover, J. E. & Strick, P. L. Multiple output channels in the basal ganglia. *Science* **259**, 819–821 (1993).
- Lynch, J. C., Hoover, J. E. & Strick, P. L. Input to the primate frontal eye field from the substantia nigra, superior colliculus and dentate nucleus demonstrated by transneuronal transport. *Exp. Brain Res.* **100**, 181–186 (1994).
- Tian, J. R. & Lynch, J. C. Functionally defined smooth and saccadic eye movement subregions in the frontal eye field of Cebus monkey. *J. Neurophysiol.* **76**, 2740–2753 (1996).
- Tian, J. R. & Lynch, J. C. Slow and saccadic eye movements evoked by microstimulation in the supplementary eye field of the cebus monkey. *J. Neurophysiol.* **74**, 2204–2210 (1995).
- Schleicher, A. et al. A stereological approach to human cortical architecture: identification and delineation of cortical areas. *J. Chem. Neuroanat.* **20**, 31–47 (2000).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank M. Bouchard for technical assistance. This work was supported by grants from Canadian Institutes of Health Research and the James S. McDonnell Foundation.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to M.P. (petrides@ego.psych.mcgill.ca).

Phosphoinositide phosphatase activity coupled to an intrinsic voltage sensor

Yoshimichi Murata^{1,2*}, Hirohide Iwasaki^{1,2,3*}, Mari Sasaki^{1,2,3*}, Kazuo Inaba⁴ & Yasushi Okamura^{1,2,3,5}

Changes in membrane potential affect ion channels and transporters, which then alter intracellular chemical conditions. Other signalling pathways coupled to membrane potential have been suggested^{1–3} but their underlying mechanisms are unknown. Here we describe a novel protein from the ascidian *Ciona intestinalis* that has a transmembrane voltage-sensing domain homologous to the S1–S4 segments of voltage-gated channels and a cytoplasmic domain similar to phosphatase and tensin homologue. This protein, named *C. intestinalis* voltage-sensor-containing phosphatase (Ci-VSP), displays channel-like ‘gating’ currents and directly translates changes in membrane potential into the turnover of phosphoinositides. The activity of the phosphoinositide phosphatase in Ci-VSP is tuned within a physiological range of membrane potential. Immunocytochemical studies show that Ci-VSP is expressed in *Ciona* sperm tail membranes, indicating a possible role in sperm function or morphology. Our data demonstrate that voltage sensing can function beyond channel proteins and thus more ubiquitously than previously realized.

During our systematic genomic survey of the ascidian *C. intestinalis* (unpublished observations), we discovered a gene that was homologous to both an ion channel and a phosphatase (ci0100145019 from JGI *Ciona* v1.0, <http://genome.jgi-psf.org/ciona4/ciona4.home.html>). By isolating its complementary DNA with reverse transcriptase polymerase chain reaction (RT-PCR) and determining its full coding sequence, we were able to deduce its likely membrane topology from the amino acid sequence⁴. Like voltage-gated ion channels, this protein contained four transmembrane segments (Fig. 1a), with four positively charged amino acids periodically aligned with two intervening hydrophobic residues in the putative S4 segment (Fig. 1b), and negative charges in the putative S2 and S3 segments. The phosphatase-like domain of this protein shared many structural features with phosphatase and tensin homologue (PTEN). For example, it contained a conserved HCXXGXXR signature motif, as required for the activity of both tyrosine phosphatase and phosphoinositide phosphatase^{5,6} (Fig. 2a), and positively charged residues that interact with the phosphate groups of the phospholipid substrate (His 332, Lys 364 and Lys 367). We therefore named this protein *C. intestinalis* voltage-sensor-containing phosphatase (Ci-VSP).

To test whether the putative transmembrane region of Ci-VSP operates as a voltage sensor, we microinjected complementary RNAs of Ci-VSP into *Xenopus* oocytes and recorded asymmetric displacement currents. Oocytes expressing Ci-VSP showed robust transient outward currents in response to depolarization and inward currents in response to hyperpolarization, resembling the ON and OFF gating currents of voltage-gated channels, respectively (Fig. 1c). We

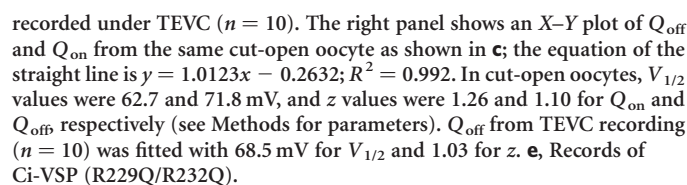
constructed a Q–V curve of the ON and OFF current fitted with the Boltzmann equation (Fig. 1d). The carried charge of the ON current coincided with that of the OFF current over a wide range of membrane voltages (Fig. 1d, right). Compared with the Shaker-type K⁺ channel⁷, the voltage-dependence of charge movement was shifted positively ($V_{1/2}$ of +62.7 mV for Ci-VSP, compared with –37.9 mV for the Shaker K⁺ channel) and the slope (represented as the value of the effective valence, z) was less steep (1.8 for z in the ON current of the Shaker K⁺ channel⁸, compared with 1.3 for z in the ON current of Ci-VSP). The channel-like ‘gating’ current of Ci-VSP resulted exclusively from the transmembrane segment, because it persisted in a protein lacking the whole carboxy-terminal cytoplasmic domain. Also like voltage-gated channels, the ‘gating’ current of Ci-VSP depended on positive charges in the S4-like segment, because it disappeared when two arginines (Arg 229 and Arg 232) among the four positively charged residues were mutated to the neutral glutamine (Fig. 1e). We confirmed the presence of this mutant protein on the cell surface by biotinylation labelling with anti-Ci-VSP antibody (data not shown).

PTEN dephosphorylates phosphatidylinositol-3,4,5-trisphosphate (PtdIns(3,4,5)P₃) at the 3-site of its inositol ring, which results in the formation of PtdIns(4,5)P₂ (ref. 9). To see whether the homologous phosphatase-like domain of Ci-VSP could also generate PtdIns(4,5)P₂ from PtdIns(3,4,5)P₃, we synthesized a fusion protein of glutathione *S*-transferase (GST) and the C-terminal cytoplasmic region in *Escherichia coli*. We first confirmed the dephosphorylating activity of Ci-VSP by TLC analysis¹⁰ in which fluorescently labelled PtdIns(3,4,5)P₃ was used as a substrate (Fig. 2b). Next we used a malachite green assay¹¹, which showed a release of phosphate significantly higher than the basal reaction containing GST protein alone (Fig. 2c). Finally, in PTEN, the cysteine residue in the catalytic domain is essential for phosphatidylinositol phosphatase activity⁹. Predictably, mutation of the catalytic cysteine residue to serine (C363S) in Ci-VSP nullified its activity (Fig. 2b, lanes 7 and 8, and Fig. 2c). Analysis of enzymatic kinetics (Fig. 2d, e) showed that the K_m against PtdIns(3,4,5)P₃ was about 36 μ M, which is consistent with the values reported for PTEN⁶.

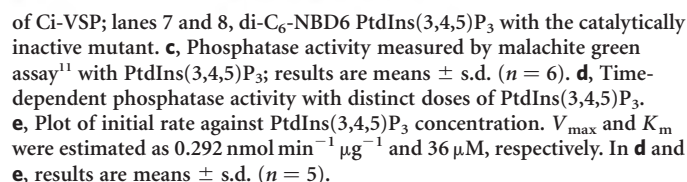
To test whether the enzymatic activity of Ci-VSP was dependent on membrane potential, we used the activities of inwardly rectifying potassium (Kir) channels and M-current K⁺ channels, which are known to be activated by phosphoinositides, mainly PtdIns(4,5)P₂ (refs 12–14). When GIRK2 (Kir3.2) channels with bovine G protein β_1 and γ_1 were expressed without Ci-VSP, the current amplitude of Kir currents activated by stepping to –100 mV did not change after repeated pulses with 10-s intervals between test pulses at either –60 mV or 0 mV. In contrast, when Ci-VSP was expressed together

¹Section of Developmental Neurophysiology, Okazaki Institute for Integrative Bioscience, National Institutes of Natural Sciences, Higashiyama 5-1, Myodaiji-cho, Okazaki, Aichi, 444-8787, Japan. ²National Institute for Physiological Sciences, National Institutes of Natural Sciences, Nishigonaka 38, Myodaiji-cho, Okazaki, Aichi, 444-8585, Japan. ³School of Life Science, The Graduate University for Advanced Studies, Okazaki 444-8585, Japan. ⁴Shimoda Marine Research Center, Graduate School of Life and Environmental Sciences, University of Tsukuba, 5-10-1 Shimoda, Shizuoka 415-0025, Japan. ⁵Neuroscience Research Institute, National Institute of Advanced Industrial Science and Technology, Tsukuba Central 6, Higashi 1-1-1, Tsukuba, Ibaraki 305-8566, Japan.

*These authors contributed equally to this work.



amplitude depended on the enzyme activity of Ci-VSP, because it was not seen in cells expressing a mutant of Ci-VSP lacking enzyme activity, namely C363S (Fig. 3a, third row of panels). In addition, Kir currents did not change when a 'gating'-current-defective mutant of Ci-VSP (Fig. 1e) was expressed (Fig. 3a, bottom row of panels),



indicating that the changes in Kir currents depended on the voltage-sensing nature of Ci-VSP. Similar results were obtained with another GIRK isoform, GIRK4(S143T) (a homomeric mutant of Kir3.4; data not shown).

GIRK channels (Kir3 subfamily) are equally sensitive to a broad spectrum of phosphoinositides¹⁵. In contrast, the constitutively active IRK1 (Kir2.1) channels have a higher specificity for PtdIns(4,5)P₂ (ref. 15). When IRK1 was expressed together with Ci-VSP, there was no change in current amplitude (Fig. 3b and Supplementary Fig. 1a). We suspected that the sensitivity of IRK1 to PtdIns(4,5)P₂ was too high to reflect the dynamic change in PtdIns(4,5)P₂ concentration¹⁴. We therefore used an IRK1 mutant in which one of the sites responsible for the PtdIns(4,5)P₂ sensitivity of IRK1, Arg 228, was replaced by a residue with reduced sensitivity to PtdIns(4,5)P₂ (ref. 14), namely glutamine (R228Q). IRK1(R228Q), like GIRK2, showed a change of current amplitude that was dependent on the potential between test pulses (Fig. 3b and Supplementary Fig. 1a). Finally, we tested KCNQ2/3 channels, which underlie M-currents and are activated by PtdIns(4,5)P₂ (ref. 13), and observed similar current responses in the presence of Ci-VSP to those of the Kir channels (Supplementary Fig. 1b). These results indicate that in

the presence of Ci-VSP, PtdIns(4,5)P₂ concentration is changed in a manner that is dependent on membrane potential (Fig. 3b).

In some cells, the current amplitudes of GIRK channels were measured at varied holding potentials. Current amplitudes were smaller at a depolarized interval potential and saturated at about 0 mV. The increase of current amplitude saturated at about -80 mV (Fig. 3c, d). A similar pattern was obtained when the holding potential was changed in the reverse sequence. In the four cells tested, the voltage-dependence profiles of GIRK currents were similar (Fig. 3d) and were shifted in the negative direction compared with the Q-V curve of voltage sensor movement (Fig. 1d; Fig. 3d, red curve). The Q-V curve did not shift significantly when the holding potential was varied. Only 10–15% of charge movements corresponded to the decrease in GIRK channel activities near saturated levels.

To test whether the linker region between the S4-like segment and the enzyme domain was critical for signal coupling in Ci-VSP, as it is between the voltage sensor and the channel pore in voltage-gated channels¹⁶, an eight-amino-acid segment was deleted in two places, 240–247 (Δ N-L) and 248–255 (Δ C-L). Both deletion constructs had gating currents indistinguishable from those of wild-type protein

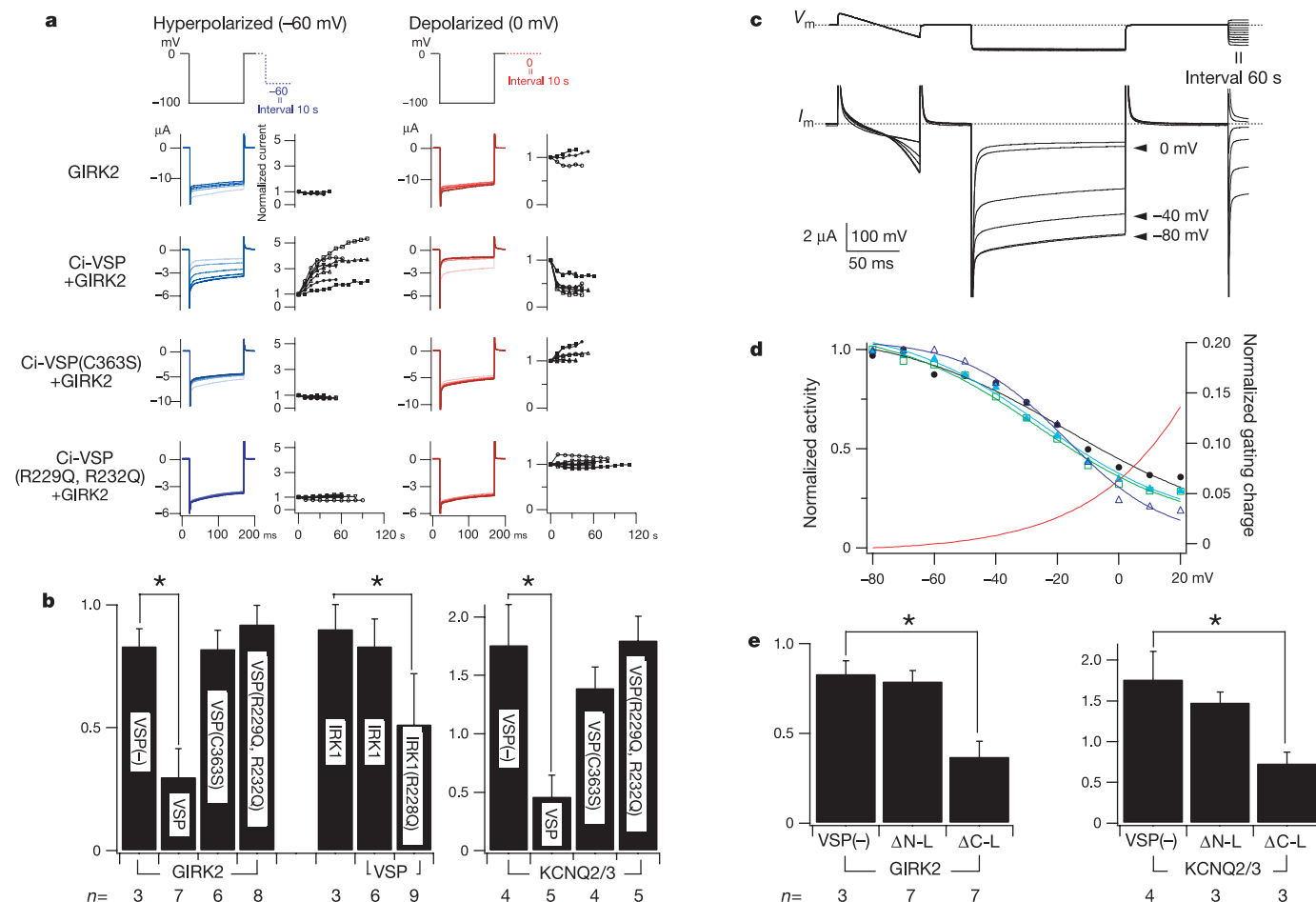


Figure 3 | Ci-VSP alters phosphoinositide concentration in a voltage-dependent manner as probed with K⁺ channel activities. **a**, Changes of Kir currents on hyperpolarization (left) or depolarization (right) during intervals between test pulses. Thicker traces were recorded later. The amplitude of Kir current during each test pulse was standardized by the amplitude obtained at zero time. **b**, Averaged ratios of the minimum currents against the maximum current from repeated test pulses (Kir) and the current at first test pulse (KCNQ2/3). Asterisk indicates a statistically significant difference ($P < 0.05$) from Ci-VSP(-) cells. **c**, Records of GIRK

currents on expression together with Ci-VSP with varied voltages during intervals test pulses ranging from -80 mV to 20 mV. **d**, Plot of current amplitudes of GIRK currents against the voltages of intervals between test pulses. Records from four different cells are shown as distinct symbols. The red line is the Q-V curve, which is identical to that shown in Fig. 1d. **e**, Summary of K⁺ channel currents with deletion of N-terminal (Δ N-L) or C-terminal (Δ C-L) half of the linker region. Error bars in **b** and **e** show s.d. for the indicated number of experiments.

(Supplementary Fig. 2c). However, $\Delta N-L$ showed no significant change in the amplitude of Kir currents by holding at either -60 mV or 0 mV for 10-s intervals between test pulses (Supplementary Fig. 2b, summarized in Fig. 3e). This effect was not due to the reduced linker size, because $\Delta C-L$ showed inter-pulse-dependent changes in Kir currents. Similar results were obtained with KCNQ2/3 channel currents (Fig. 3e, right panel). A short stretch flanking the S4-like segment is therefore important in coupling between the two domains.

RT-PCR showed that Ci-VSP is expressed abundantly in testis and weakly in neural complex (Fig. 4a). Expression in sperm was further examined by using polyclonal antibodies raised against the amino-terminal peptide of Ci-VSP (Supplementary Fig. 3). Positive signals were detected on the sperm tail (Fig. 4b–d). Signals in the sperm head, probably in mitochondria, were nonspecific, because they remained when antibody was competed with antigen peptides (Fig. 4e–g, and insets). Similar results were obtained with antibody

against the C-terminal peptide of Ci-VSP (data not shown). By immunoelectron microscopic observation (Fig. 4h–k), signals of gold particles were detected in the putative membrane regions (arrows in Fig. 4h–k).

Changes in phosphatase activity of Ci-VSP reported by K^+ channel activities are already saturated at about $+20$ mV, where only less than 15% of the charge movements of the voltage sensor occurs (Fig. 3d). This contrasts with voltage-gated channels, in which charge movement of the voltage sensor is saturated at less positive membrane potentials than the curve of the probability of channel opening¹⁷. Such a large shift in the voltage dependence of predicted enzyme activity against movement of the voltage sensor could be due to a possible nonlinearity in the sensitivity of GIRK channel activities to phosphoinositide in our measurements. Alternatively, a structural change associated with early events of charge movement could be sufficient to induce a conformational change in the phosphatase domain.

Ci-VSP is the first example of a non-channel protein whose activity is regulated by membrane potential through an S1–S4-based voltage sensor. How can movement of the voltage sensor lead to modification of the enzyme domain, which has a cytoplasmic globular structure? The N-terminal end of PTEN is known to be critical for phosphatase activity^{18,19}. The S4-like segment of Ci-VSP might induce movement of the N-terminal end, which then triggers a conformational change of the phosphatase domain such as a change in the spatial relationship between catalytic centre and lipid substrate. Further characterization of Ci-VSP will lead to a better understanding of the general molecular principles of coupling between a voltage sensor and its effector.

Expression of Ci-VSP in sperm suggests that it has a function in sperm physiology²⁰ or morphology. In *C. intestinalis*, sperm-activating and attracting factor, a recently identified steroid²¹, is known to hyperpolarize sperm membrane, leading to the activation of sperm motility²⁰. Through its phosphatase activity, Ci-VSP could mediate the regulation of sperm motility after hyperpolarization. TPTE/PTEN2/TPIP, a mammalian VSP homologue gene, is expressed in testis^{22,23}, suggesting a conserved functional role of VSP in testis among chordate species. TPTE/PTEN2/TPIP has also been reported to be transiently expressed in brain and spinal cord during mouse embryogenesis²⁴ and ectopically expressed in human cancer cells²⁵. Further characterization of Ci-VSP and its vertebrate homologues will unravel previously unaddressed biological roles of membrane potential.

METHODS

Plasmid construction. For cloning of full-length Ci-VSP cDNA, total RNA was purified from young adults of *C. intestinalis* and cDNA was amplified by PCR using primers based on JGI database²⁶. The PCR products were cloned into a pBluescript vector (Stratagene) by a *Bam*HI site and a *Not*I site. The point mutants and deletion mutants were made with a QuikChange site-directed mutagenesis kit (Stratagene). For expression in *Xenopus* oocyte, cDNA clones were constructed in expression vectors, pSD64TR, pSD64TF (a gift from T. Snutch) or pGEM HE vector. cRNA was transcribed *in vitro* with an mMESAGE mMACHINE kit (Ambion).

Electrophysiology in *Xenopus* oocytes. Oocytes were surgically isolated from *Xenopus laevis*, which had first been anaesthetized by soaking in 0.15% tricaine solution. cRNA of GIRK2, G β and G γ were injected at a concentration ratio of 0.3–0.5 to that of Ci-VSP. IRK1 and IRK1(R228Q) were injected at a concentration ratio of 0.1 to Ci-VSP. The oocytes were incubated in ND96 solution²⁷ for 2–4 days at $17 \pm 1^\circ\text{C}$. Cut-open voltage-clamp recording was performed with CA-1B (DAGAN) as an amplifier²⁸. Linear and symmetrical current was subtracted by a $P/-8$ procedure (inset in Fig. 1c). The $Q-V$ curve was fitted by a Boltzmann relation, $Q = 1/[1 + \exp\{ze(V - V_{1/2})/kT\}]$, where k is the Boltzmann constant and e the elementary electric charge. Stimulation and data acquisition were performed with a Digidata 1322A AD/DA converter with Clampex software (Axon Instruments Co.). The external solution contained (in mM) 105 *N*-methyl-D-glucamine (NMDG)-methanesulphonate, 2 CaCl_2 , 10 HEPES pH 7.5. The internal solution contained (in mM) 105 NMDG-methanesulphonate, 2 MgCl_2 , 0.1 EGTA, 10 HEPES pH 7.5. Internal perfusion

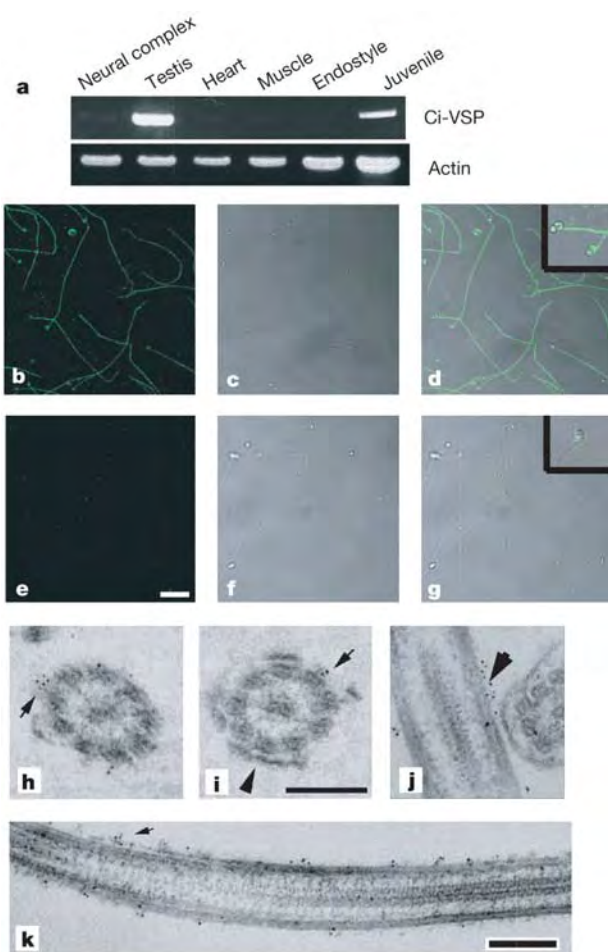


Figure 4 | Ci-VSP protein is expressed in sperm tail. **a**, Tissue expression pattern of Ci-VSP transcript detected by RT-PCR. **b–g**, Immunocytochemistry of *C. intestinalis* sperm with polyclonal antibody against Ci-VSP, showing the fluorescence signal (**b**, **e**), phase-contrast images (**c**, **f**) and merged views (**d**, **g**). **b–d**, Fluorescence was observed in the flagella and mitochondria. **e–g**, As a negative control, the antibody preabsorbed with antigenic peptide was used. Signals in the sperm head still remained, suggesting that they are nonspecific. Scale bar, $10\ \mu\text{m}$. **h–k**, Immunoelectron localization of Ci-VSP at plasma membrane of sperm flagella. Example cross sections (**h**, **i**) and longitudinal sections (**j**, **k**) of the flagella are shown. The peripheral regions of the flagellar axonemes were labelled where debris of plasma membrane are associated (arrows). Gold particles were found only on the debris of partly disrupted plasma membrane (arrows), but not on apparently intact plasma membrane (arrowhead in **i**). Scale bar, $200\ \text{nm}$ (the scale bar in **i** also applies to **h** and **j**).

was achieved through a glass pipette at $2\ \mu\text{L min}^{-1}$ with a syringe pump.

For measuring Kir currents or KCNQ2/3 currents, two-electrode voltage-clamp recording (TEVC) was performed with a 'bath-clamp' amplifier OC-725C. The microelectrode resistance ranged from 0.1 to 0.6 M Ω . The standard bath solution contained (in mM) 92 KCl, 3 MgCl₂, 4 KOH, 5 HEPES pH 7.4 for Kir channel recordings and 2 KCl, 92 NaCl, 3 MgCl₂, 4 NaOH, 5 HEPES pH 7.4 for measurements of KCNQ2/3 channel currents. All of GIRK cRNAs were co-injected with G protein β_1 and γ_1 . Changes in the leakage current of the Kir current were monitored by applying an 80-ms ramp pulse before the test pulse. Data from cells showing a leakage current of more than 1 μA at a holding potential of 0 mV or during the ramp pulse were discarded. GIRK currents were activated by stepping to $-100\ \text{mV}$ for 150 ms with a 10-s interval at either $-60\ \text{mV}$ (left) or 0 mV (right). Output current was filtered by a four-pole Bessel filter at 1 kHz. Sampling frequency was 13–27.7 kHz for TEVC recording and 50–100 kHz for cut-open oocytes. All electrical recordings were performed at 23–27°C. Data were analysed with Igor Pro (WaveMetrics Inc.). Statistical significance was determined as $P < 0.05$ with a Mann–Whitney U-test. Results are presented as means $\pm$ s.d.

In vitro phosphatase assay. To express N-terminal fused protein of the cytoplasmic phosphatase domain of Ci-VSP with the GST gene, the cDNA encoding residues 248–576 of Ci-VSP was amplified by PCR and subcloned into the *EcoRI*–*XhoI* site of pGEX4T3 (Amersham Pharmacia). GST-fused protein was expressed and purified from *E. coli* JM109 with glutathione–Sephadex 4B (Amersham Pharmacia). Phosphatase activity was measured with the methods previously adopted for PTEN phosphatase^{10,11} (Supplementary Methods).

Immunocytochemistry. Rabbit polyclonal anti-Ci-VSP antibody was raised against a synthetic oligopeptide corresponding to amino-acid residues 85–96 (ENEHGVDGGRME) of Ci-VSP. Sperm were fixed with 4% paraformaldehyde, permeabilized with 0.1% Triton X-100 and incubated with the anti-Ci-VSP antibody (2.5 $\mu\text{g ml}^{-1}$ final concentration) and then with the secondary antibody Alexa 488-conjugated anti-rabbit IgG (Molecular Probes). As a control, the antigenic oligopeptide was used for the absorption of the antibody. Specimens were observed by a laser-scanning microscope (LSM510; Zeiss). For immunogold labelling²⁹, sperm were fixed with glutaraldehyde and incubated with affinity-purified anti-VSP antibody and with goat anti-rabbit secondary antibody conjugated with gold (BioCell). Samples were postfixed, dehydrated, embedded and thin-sectioned.

Received 15 September 2004; accepted 21 April 2005.

Published online 18 May 2005.

- Zhang, C. & Zhou, Z. Ca^{2+} -independent but voltage-dependent secretion in mammalian dorsal root ganglion neurons. *Nature Neurosci.* **5**, 425–430 (2002).
- Zhang, C. *et al.* Calcium- and dynamin-independent endocytosis in dorsal root ganglion neurons. *Neuron* **42**, 225–236 (2004).
- Schultz, J. E., Klumpp, S., Benz, R., Schurhoff-Goeters, W. J. & Schmid, A. Regulation of adenylyl cyclase from *Paramecium* by an intrinsic potassium conductance. *Science* **255**, 600–603 (1992).
- Hirokawa, T., Boon-Chieng, S. & Mitaku, S. SOSUI: classification and secondary structure prediction system for membrane proteins. *Bioinformatics* **14**, 378–379 (1998).
- Lee, J. O. *et al.* Crystal structure of the PTEN tumor suppressor: implications for its phosphoinositide phosphatase activity and membrane association. *Cell* **99**, 323–334 (1999).
- Maehama, T., Taylor, G. S. & Dixon, J. E. PTEN and myotubularin: novel phosphoinositide phosphatases. *Annu. Rev. Biochem.* **70**, 247–279 (2001).
- Bezanilla, F. Voltage sensor movements. *J. Gen. Physiol.* **120**, 465–473 (2002).
- Bezanilla, F., Perozo, E., Papazian, D. M. & Stefani, E. Molecular basis of gating charge immobilization in Shaker potassium channels. *Science* **254**, 679–683 (1991).
- Maehama, T. & Dixon, J. E. The tumor suppressor, PTEN/MMAC1, dephosphorylates the lipid second messenger, phosphatidylinositol 3,4,5-trisphosphate. *J. Biol. Chem.* **273**, 13375–13378 (1998).
- Taylor, G. S. & Dixon, J. E. PTEN and myotubularins: families of phosphoinositide phosphatases. *Methods Enzymol.* **366**, 43–56 (2003).
- Maehama, T., Taylor, G. S., Slama, J. T. & Dixon, J. E. A sensitive assay for phosphoinositide phosphatases. *Anal. Biochem.* **279**, 248–250 (2000).
- Huang, C. L., Feng, S. & Hilgemann, D. W. Direct activation of inward rectifier potassium channels by PIP₂ and its stabilization by G $\beta\gamma$. *Nature* **391**, 803–806 (1998).
- Zhang, H. *et al.* PIP₂ activates KCNQ channels, and its hydrolysis underlies receptor-mediated inhibition of M currents. *Neuron* **37**, 963–975 (2003).
- Zhang, H., He, C., Yan, X., Mirshahi, T. & Logothetis, D. E. Activation of inwardly rectifying K⁺ channels by distinct PtdIns(4,5)P₂ interactions. *Nature Cell Biol.* **1**, 183–188 (1999).
- Rohacs, T., Chen, J., Prestwich, G. D. & Logothetis, D. E. Distinct specificities of inwardly rectifying K⁺ channels for phosphoinositides. *J. Biol. Chem.* **274**, 36065–36072 (1999).
- Horn, R. Coupled movements in voltage-gated ion channels. *J. Gen. Physiol.* **120**, 449–453 (2002).
- Bezanilla, F. The voltage sensor in voltage-dependent ion channels. *Physiol. Rev.* **80**, 555–592 (2000).
- Campbell, R. B., Liu, F. & Ross, A. H. Allosteric activation of PTEN phosphatase by phosphatidylinositol 4,5-bisphosphate. *J. Biol. Chem.* **278**, 33617–33620 (2003).
- Iijima, M., Huang, Y. E., Luo, H. R., Vazquez, F. & Devreotes, P. N. Novel mechanism of PTEN regulation by its phosphatidylinositol 4,5-bisphosphate binding motif is critical for chemotaxis. *J. Biol. Chem.* **279**, 16606–16613 (2004).
- Izumi, H., Marian, T., Inaba, K., Oka, Y. & Morisawa, M. Membrane hyperpolarization by sperm-activating and -attracting factor increases cAMP level and activates sperm motility in the ascidian *Ciona intestinalis*. *Dev. Biol.* **213**, 246–256 (1999).
- Yoshida, M., Murata, M., Inaba, K. & Morisawa, M. A chemoattractant for ascidian spermatozoa is a sulfated steroid. *Proc. Natl Acad. Sci. USA* **99**, 14831–14836 (2002).
- Walker, S. M., Downes, C. P. & Leslie, N. R. TPPI: a novel phosphoinositide 3-phosphatase. *Biochem. J.* **360**, 277–283 (2001).
- Tapparel, C. *et al.* The TPTE gene family: cellular expression, subcellular localization and alternative splicing. *Gene* **323**, 189–199 (2003).
- Reymond, A. *et al.* Human chromosome 21 gene expression atlas in the mouse. *Nature* **420**, 582–586 (2002).
- Dong, X. Y. *et al.* Identification of two novel CT antigens and their capacity to elicit antibody response in hepatocellular carcinoma patients. *Br. J. Cancer* **89**, 291–297 (2003).
- Dehal, P. *et al.* The draft genome of *Ciona intestinalis*: insights into chordate and vertebrate origins. *Science* **298**, 2157–2167 (2002).
- Goldin, A. L. Maintenance of *Xenopus laevis* and oocyte injection. *Methods Enzymol.* **207**, 266–279 (1992).
- Tagliatela, M., Toro, L. & Stefani, E. Novel voltage clamp to record small, fast currents from ion channels expressed in *Xenopus* oocytes. *Biophys. J.* **61**, 78–82 (1992).
- Padma, P. *et al.* Identification of a novel leucine-rich repeat protein as a component of flagellar radial spoke in the ascidian *Ciona intestinalis*. *Mol. Biol. Cell* **14**, 774–785 (2003).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank D. E. Logothetis for providing IRK1 R228Q plasmid, Y. Kubo for providing wild-type IRK1 plasmid, T. Nukada for G-protein β_1 and γ_1 subunit plasmids, D. McKinnon and K. Nakajo for KCNQ2/3 plasmids, M. Lazdunski for GIRK2 plasmid, T. Maehama for advice on measuring phosphatase activity, F. Kukita for help in cut-open oocyte recording, J. Cui for discussion, N. Satoh and Y. Satou for help in bioinformatics, and D. McLean for critical reading of the manuscript. This work was supported by Research Fellowships of the Japan Society for the Promotion of Science for Young Scientists to Y.M., Grants-in-Aid for Scientific Research to Y.O. and H.I., and a Grant-in-Aid for Creative Scientific Research from the Ministry of Education, Culture, Sports, Science and Technology to Y.O.

Author Information The full-length cDNA sequence of Ci-VSP is deposited in the DNA Data Bank of Japan (DDBJ) under the accession number AB183035. Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to Y.O. (yokamura@nips.ac.jp).

LETTERS

Ephrin signalling controls brain size by regulating apoptosis of neural progenitors

Vanessa Depaeppe¹, Nathalie Suarez-Gonzalez¹, Audrey Dufour¹, Lara Passante¹, Jessica A Gorski², Kevin R. Jones², Catherine Ledent¹ & Pierre Vanderhaeghen¹

Mechanisms controlling brain size include the regulation of neural progenitor cell proliferation, differentiation, survival and migration^{1,2}. Here we show that ephrin-A/EphA receptor signalling plays a key role in controlling the size of the mouse cerebral cortex by regulating cortical progenitor cell apoptosis. *In vivo* gain of EphA receptor function, achieved through ectopic expression of ephrin-A5 in early cortical progenitors expressing EphA7, caused a transient wave of neural progenitor cell apoptosis, resulting in premature depletion of progenitors and a subsequent dramatic decrease in cortical size. *In vitro* treatment with soluble ephrin-A ligands similarly induced the rapid death of cultured dissociated cortical progenitors in a caspase-3-dependent manner, thereby confirming a direct effect of ephrin/Eph signalling on apoptotic cascades. Conversely, *in vivo* loss of EphA function, achieved through EphA7 gene disruption, caused a reduction in apoptosis occurring normally in forebrain neural progenitors, resulting in an increase in cortical size and, in extreme cases, exencephalic forebrain overgrowth. Together, these results identify ephrin/Eph signalling as a physiological trigger for apoptosis that can alter brain size and shape by regulating the number of neural progenitors.

Ephrin and Eph receptor genes have been implicated in the control of cell and axon guidance in many neural systems^{3–5}. To identify additional roles for ephrin/Eph genes in the developing forebrain, we generated a mouse transgenic line that allows ectopic expression of ephrin-A5 under the control of the EphA7 receptor promoter, taking advantage of their complementary patterns of expression. Ephrin-A5 and EphA7 are expressed as early as embryonic day (E)10.5 in the forebrain. EphA7 is preferentially expressed in cortical progenitors, and ephrin-A5 is expressed in progenitors of the ventral telencephalon and the medial dorsal telencephalon (Fig. 1a). Perinatally, they are expressed in complementary gradients in the cerebral cortex (Fig. 1a and ref. 6). To faithfully reproduce the EphA7 expression pattern of the ephrin-A5 transgene without interfering with EphA7 function, we used a transgenic bacterial artificial chromosome (BAC) approach, in which a conditional, Cre-recombinase-dependent, lox/enhanced green fluorescent protein (eGFP)/lox/ephrin-A5 expression cassette was placed under the control of EphA7 regulatory elements using BAC recombineering^{7,8} (Fig. 1b). The resulting transgenic mice (TGA7A5) expressed the conditional cassette in a pattern very similar to the endogenous pattern of EphA7, both at early and later stages of cortical development, as assessed by expression of the eGFP transgene (Fig. 1c, d). No expression of the Flag-ephrin-A5 transgene was detected in the absence of Cre-recombinase. To drive expression of the ephrin-A5 transgene specifically in the developing cortex, we crossed this line with an Emx1^{IR^{ES}Cre} line (Emx1-Cre) that promotes efficient and specific

Cre-recombination in the dorsal telencephalon from E10.5 (ref. 9).

Although TGA7A5 mice and Emx1-Cre mice did not show any overt phenotype (ref. 9 and data not shown), mice doubly transgenic for TGA7A5 and Emx1-Cre (TG/Emx1-Cre) died at birth and showed a severe reduction in cerebral hemisphere size (up to threefold by E16.5; Fig. 2a), with the rest of the brain appearing normal (Fig. 2a, b). This reduction in cortical size was clearly visible as early as E12.5 (Fig. 2a, b), and was accompanied by a decrease in the tangential extent and radial thickness of the post-mitotic cortical plate at later stages (Fig. 2b), suggesting that the phenotype resulted from an impairment in early neurogenesis in the dorsal forebrain.

Such a phenotype could result from dysregulation of cell proliferation, migration, differentiation or death. Staining for nestin revealed no decrease in the density of cortical progenitor cells (Fig. 2c). To detect abnormal patterns of proliferation or cell cycle progression among cortical progenitors, we used 5-bromodeoxyuridine (BrdU) pulse-labellings combined with Ki67 staining (to mark proliferating cells). Both long and short time courses of BrdU/Ki67 labelling showed a normal labelling index (the proportion of BrdU-labelled cells among Ki67⁺ progenitors) for cortical progenitors, suggesting that their proliferative pattern and cell cycle were unimpaired in the mutant, even though their total number was reduced (Fig. 2c, e). Next, we measured the fraction of cells leaving the mitotic cycle (the quitting fraction, or the proportion of Ki67⁺ post-mitotic cells among BrdU-labelled cells)¹⁰. No difference was observed between the controls and mutants, suggesting that cortical neurogenesis *per se* was occurring normally in the mutants (Fig. 2e). In addition, analysis of pan-neural (MAP-2 and Tuj1) and specific (reelin, calretinin, Tbr1, and Brn1) markers showed a normal proportion and position of neurons and progenitors at all stages examined (see Fig. 2c, d and Supplementary Fig. 1), confirming that neural differentiation and migration appeared to take place normally in TG/Emx1-Cre mutants.

We next looked for dysregulation of cell death in the mutants (Fig. 3). A large increase in the number of dying early cortical progenitors was observed at E11.5 and E12.5, as labelled by terminal deoxynucleotidyltransferase-mediated dUTP nick end labelling (TUNEL) (Fig. 3a, b). A similar increase was observed for cells immunoreactive for activated caspase-3, one of the main effectors of developmental apoptotic death^{11,12} (Fig. 3a, b). Co-labelling for these apoptosis markers and nestin showed that the bulk (more than 90%) of dying cells were cortical progenitors (Fig. 3c). This apoptosis was restricted both spatially and temporally in a pattern similar to that observed for EphA7 and the ephrin-A5 transgene⁶ (Figs 1a, c, d and 3b). Specifically, apoptosis was restricted to the developing cerebral cortex (Fig. 3a, b) and was transient, peaking at E11.5–E12.5, reduced by E14.5 and essentially absent by E17.5 (Fig. 3d, e), except for the

¹Institut de Recherches Interdisciplinaires en Biologie Humaine et Moléculaire (IRIBHM), University of Brussels, Campus Erasme, 808 Route de Lennik, B-1070 Brussels, Belgium.

²Department of Molecular, Cellular, and Developmental Biology, University of Colorado, Boulder, Colorado 80309, USA.

activation of caspase-3 in a few post-mitotic neurons in the cortical plate (Fig. 3d).

The rapid and transient nature of the induction of apoptosis of cortical progenitors upon ectopic expression of ephrin-A5, together with the lack of evidence for significant disruption of cell migration or differentiation, strongly indicated a direct pro-apoptotic effect of ephrin/Eph signalling in these cells. To test this hypothesis, we turned to an *in vitro* assay in which cultured cortical progenitors (at E12–E13, a stage at which most cortical progenitors express EphA7 (data not shown)) were challenged with EphA agonists (consisting of clustered ephrin-A5–Fc fragment fusion proteins). Ephrin stimulation of early cortical progenitors resulted in a rapid induction of cell death, as measured by TUNEL staining (Fig. 3f). This cell death was inhibited by the pan-caspase inhibitor z-VAD-fmk (Fig. 3f), and was concurrent with an increase in caspase-3 activity (Fig. 3g), indicating that ephrins can directly trigger cortical

progenitor cell death through a caspase-dependent pathway.

These gain-of-function data suggested a potentially important function for ephrin/Eph signalling in the regulation of progenitor cell death in the developing forebrain. To test for a physiological role of Eph signalling in this context, we next turned to a loss-of-function analysis, using EphA7 knockout (EphA7^{-/-}) mice¹³. We found an approximately twofold decrease in cortical progenitor cell death at mid-neurogenesis (E13.5, which corresponds to a peak of apoptosis during cortical neurogenesis^{14,15}), as assessed by quantification of TUNEL-positive and activated caspase-3-positive cortical progenitors, double-stained for nestin (Fig. 4a–c). Notably, analysis of the labelling index, quitting fraction, and the expression of specific neural markers at the same stage of corticogenesis indicated that neural proliferation and differentiation proceeded normally in EphA7^{-/-} mutants (see Fig. 4g and Supplementary Fig. 2).

In addition, EphA7^{-/-} mutants showed a significant increase in

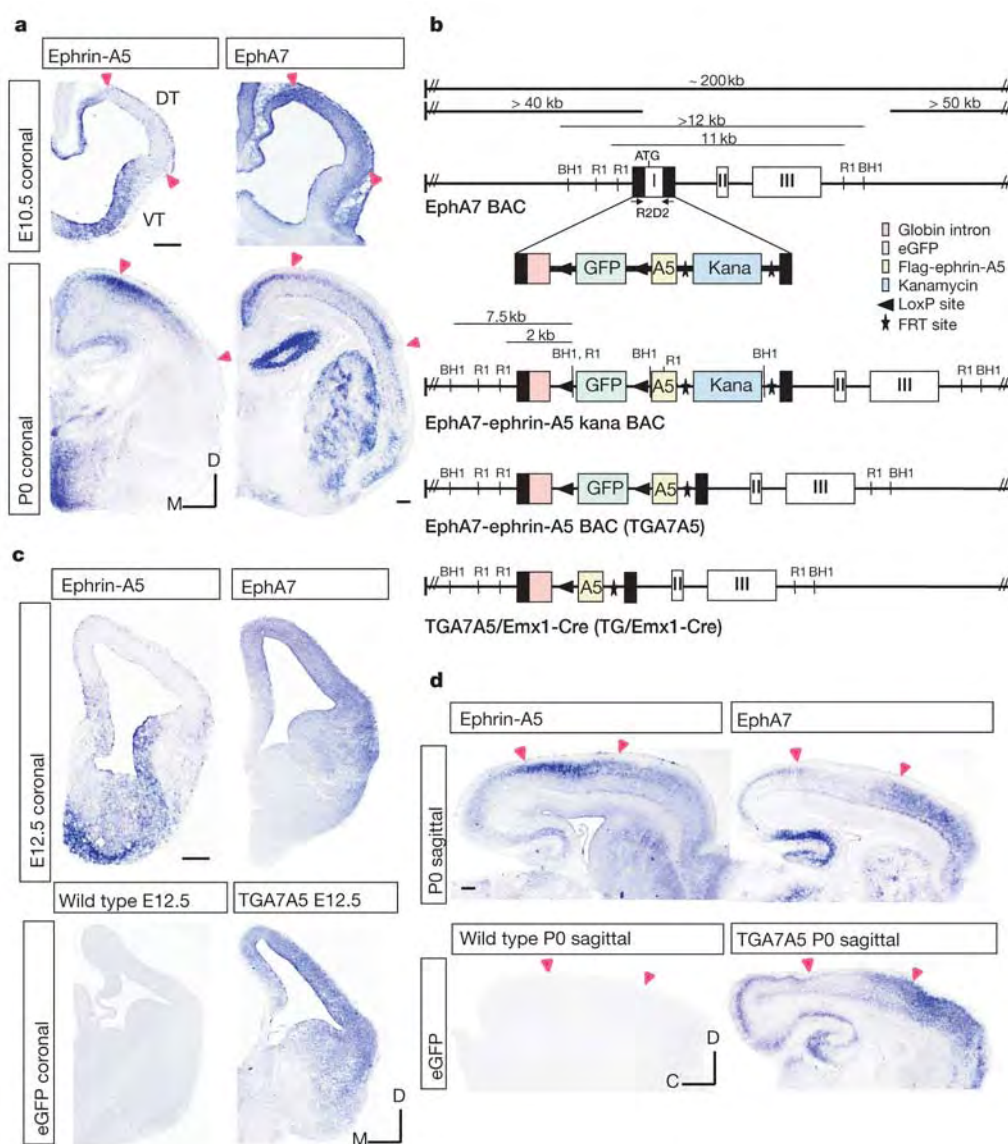


Figure 1 | Generation of EphA7/ephrin-A5 transgenic mice. **a**, *In situ* hybridization (ISH) on forebrain sections shows complementary expression patterns of ephrin-A5 and EphA7 at E10.5 and P0. DT, dorsal telencephalon; VT, ventral telencephalon. Arrowheads show the approximate boundaries of the neocortex. **b**, A BAC centred on EphA7 exon-1 was targeted with a conditional expression cassette to generate a transgenic mouse line (TGA7A5) allowing expression of eGFP under the control of EphA7

regulatory sequences. After crossing this line with Emx1-Cre mice (TG/Emx1-Cre) the eGFP cassette is removed by Cre recombinase, allowing expression of ephrin-A5 in the DT. **c**, **d**, ISH on E12.5 (**c**) and P0 (**d**) brain sections illustrates the distribution of ephrin-A5, EphA7 and the eGFP transgene. Arrowheads indicate boundaries of expression for ephrin-A5 versus EphA7/eGFP. Medial (M) is to the left (**a**, **c**), caudal is to the left (**d**), dorsal (D) is at the top in (**a**, **c**, **d**). Scale bar, 200 μ m.

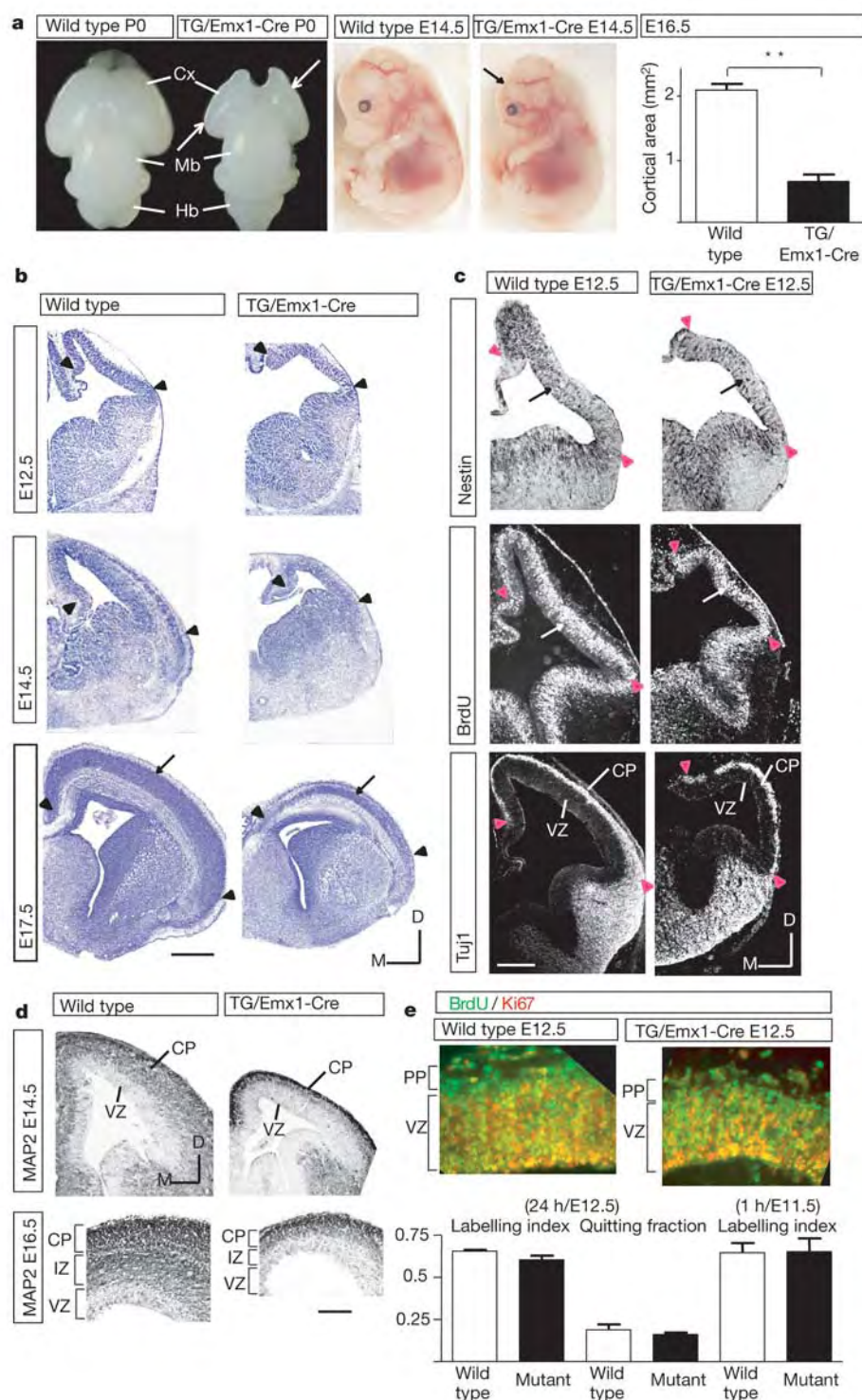


Figure 2 | Reduction in cortical size in TG/Emx1-Cre mutants. **a**, From E14.5 to P0 the size of the cortex (Cx, arrows) of TG/Emx1-Cre mice is severely reduced, but the rest of the brain (Mb, midbrain; Hb, hindbrain) is normal in size. Quantification of cortical size at E16.5 reveals a threefold reduction in TG/Emx1-Cre mice (results represent mean cortical area $\pm$ s.e.m.; $P = 0.0022$, $n = 3$ per genotype). **b**, Coronal brain sections stained with cresyl violet, illustrating the reduction of the cortical plate in TG/Emx1-Cre mice (arrows and arrowheads). Scale bar, 500 μ m. **c**, **d**, Coronal brain sections of mice at E12.5 (**c**), E14.5 and E16.5 (**d**) stained with nestin, BrdU, Tuj1 and MAP2. Despite the size reduction (red

arrowheads), the overall cellular organization in the cortical plate (CP), intermediate zone (IZ) and ventricular zone (VZ) is normal in TG/Emx1-Cre mutants. Arrows in **c** indicate nestin- and BrdU-stained cortical cells. Scale bar in **c** and **d**, 250 μ m. **e**, BrdU labelling (green) and Ki67 staining (red) reveal normal labelling index and quitting fraction in the cortex of TG/Emx1-Cre mutants. PP, cortical preplate. Results represent the mean number $\pm$ s.e.m. of counted cells (detected by BrdU or Ki67) from three animals per genotype. $P = 0.2$ and 0.9 for labelling index after 24 h and 1 h, respectively; $P = 0.4$ for quitting fraction.

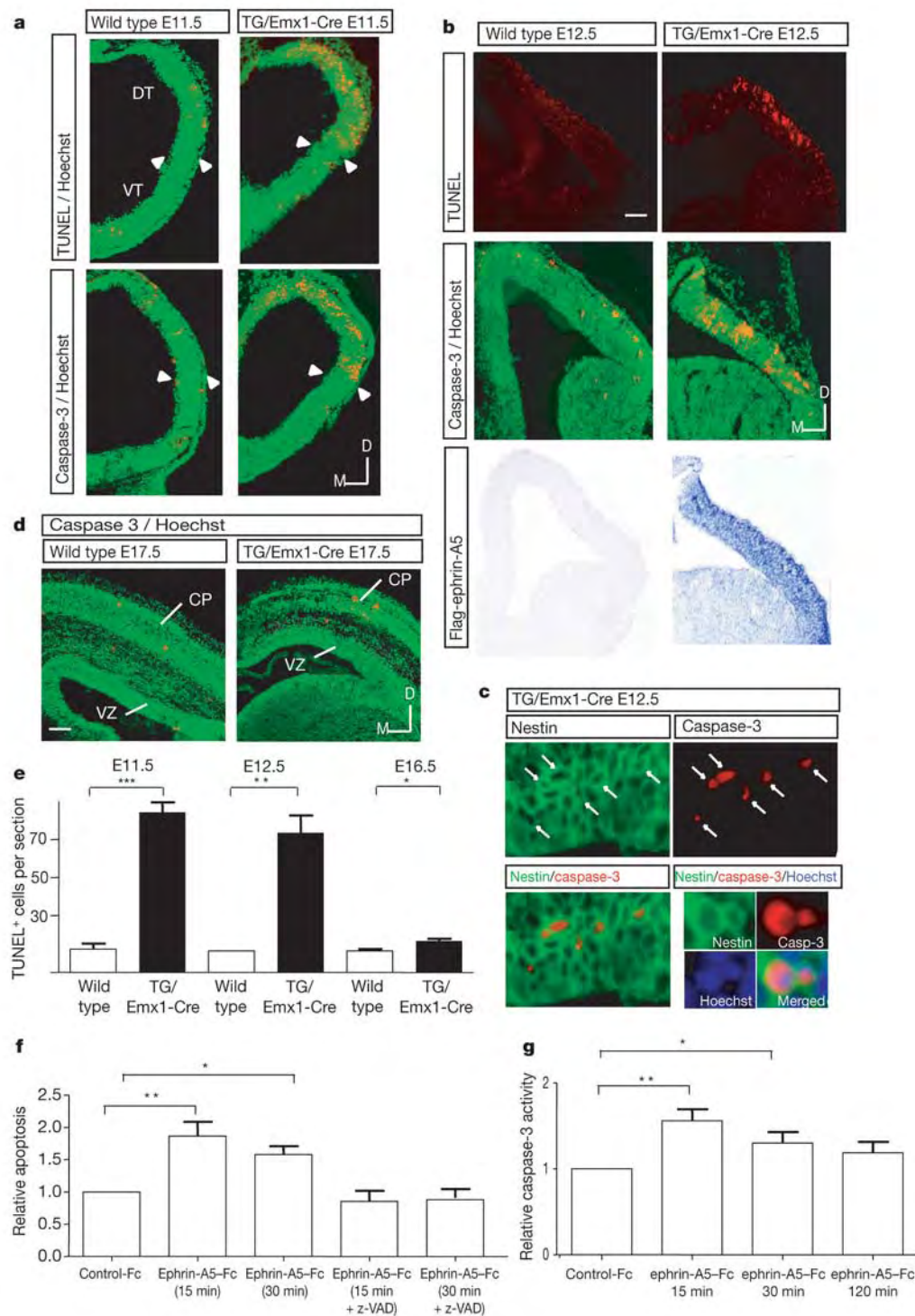


Figure 3 | Ephrin-A/EphA stimulation triggers cortical progenitor cell apoptosis *in vivo* and *in vitro*. **a, b**, At E11.5 (**a**) and E12.5 (**b**), the DT of TG/Emx1-Cre animals shows an increase in the number of apoptotic cells, as labelled by TUNEL (red) or immunostained for activated caspase-3 (red; Hoechst stain in green). Levels of apoptotic cells correlate with the expression of the ephrin-A5 transgene (Flag-ephrin-A5, detected by ISH). Arrowheads indicate the limits of DT versus VT. **c**, Dual labelling of TG/Emx1-Cre cortex with nestin (green) and caspase-3 (red) shows that most apoptotic cells are nestin⁺ (white arrows). **d**, At later stages (E17.5), no increase in apoptosis is detected in the VZ, but a few apoptotic cells are found in the CP. Scale bar in **b** and **d**, 125 μ m. **e**, Quantification of TUNEL⁺ cells in the cortex. Results represent mean number $\pm$ s.e.m. of TUNEL⁺ cells normalized to the number of brain sections used ($n = 3$ animals per

genotype). Single asterisk, $P = 0.02$; two asterisks, $P = 0.005$; three asterisks, $P < 0.0001$. **f**, Treatment of dissociated cortical progenitors with clustered ephrin-A5-Fc (100 nM) induces a rapid increase in cell death, which is inhibited by the pan-caspase inhibitor z-VAD-fmk. Results represent the mean value $\pm$ s.e.m. from six independent experiments performed in duplicate. Single asterisk, $P = 0.0112$ ($n = 6$); two asterisks, $P = 0.0065$ ($n = 6$). **g**, Clustered ephrin-A5-Fc (100 nM) also induces a rapid increase in caspase-3 activity. Single asterisk, $P = 0.0411$ ($n = 11$); two asterisks, $P = 0.0021$ ($n = 11$). Results represent the mean value $\pm$ s.e.m. from 11 independent experiments performed in duplicate. All values are normalized to apoptosis under control conditions (100 nM clustered Fc fragments).

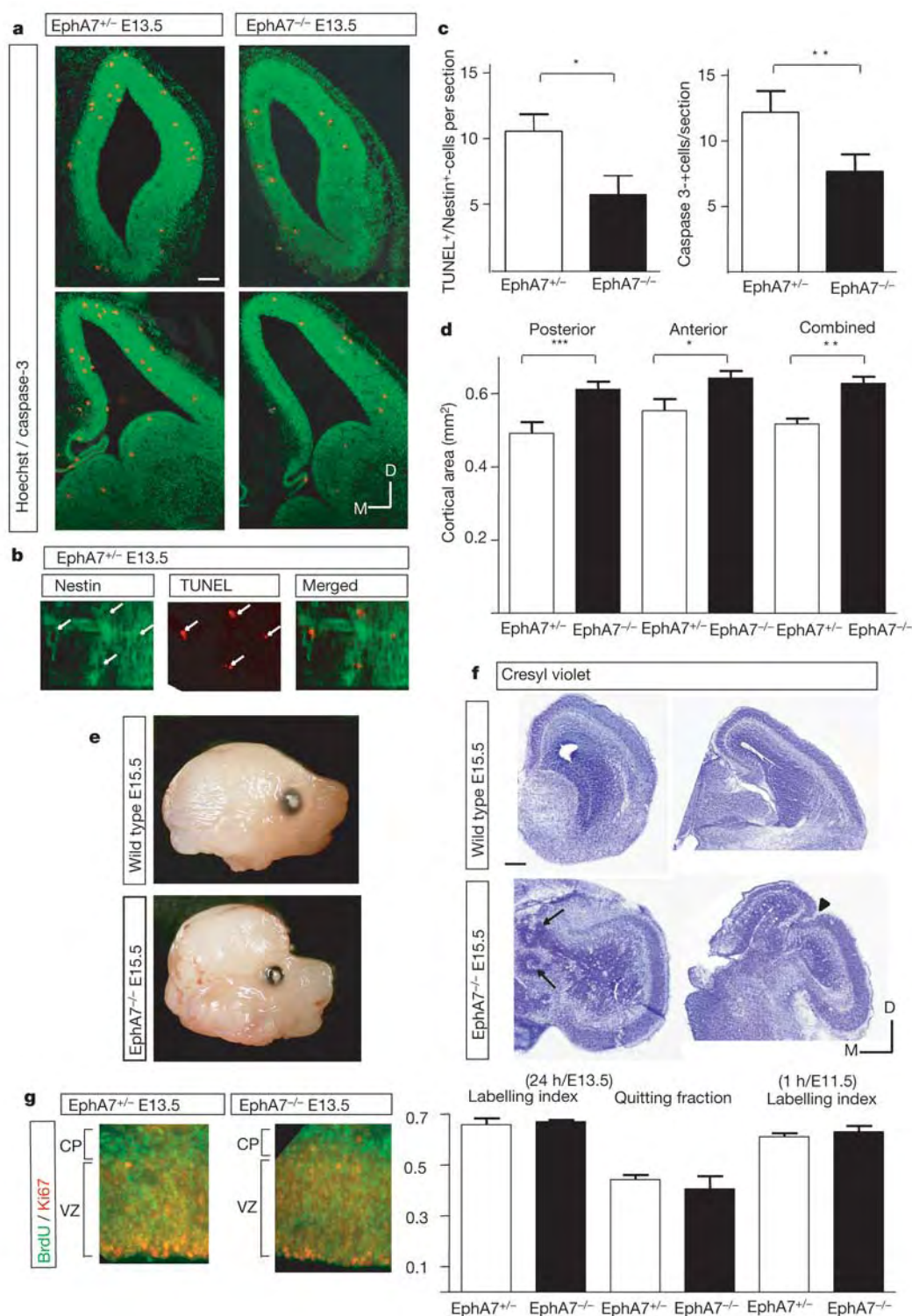


Figure 4 | Decrease in cortical progenitor cell apoptosis and increase in forebrain size in EphA7^{-/-} mutants. **a**, Coronal brain sections immunostained for activated caspase-3 (red, Hoechst in green) illustrate the decrease in apoptotic cell numbers in mutant cortex. Scale bar, 125 μ m. **b**, Co-staining with nestin (green) and TUNEL (red) shows that most apoptotic cells correspond to nestin⁺ cortical progenitors (white arrows). **c**, Quantification of cell death in E13.5 EphA7^{-/-} and EphA7^{+/-} littermates. Results show mean values $\pm$ s.e.m. of counted cells normalized to the number of brain sections counted ($n = 5$ –7 animals of each genotype). Single asterisk, $P = 0.030$ ($n = 5$ per genotype); two asterisks, $P = 0.031$ ($n = 7$ per genotype). **d**, Results represent the mean value $\pm$ s.e.m. of the

cortical area at E13.5 in EphA7^{-/-} and EphA7^{+/-} littermates, measured at two rostro-caudal levels (posterior and anterior). Single asterisk, $P = 0.030$; two asterisks, $P = 0.004$; three asterisks, $P = 0.006$ ($n = 7$ per genotype). **e**, **f**, Forebrain overgrowth in some EphA7^{-/-} mice, with ectopic proliferation (arrows in **f**) and cortical foldings (arrowhead in **f**). Scale bar, 250 μ m. **g**, BrdU labelling (green, performed after 1 h at E11.5 or 24 h at E13.5) and Ki67 staining (red). Labelling index and quitting fraction were normal in the cortex of EphA7^{-/-} mutants. Results represent the mean value $\pm$ s.e.m. of counted cells ($n = 5$ animals per genotype). $P = 0.3$ and 0.47 for the labelling index after 24 h and 1 h, respectively; $P = 0.4$ for the quitting fraction.

cortical size (~20%, at different rostro-caudal levels at E13.5) (Fig. 4d). Moreover, in a small proportion of cases (10%), EphA7^{-/-} embryos showed exencephalic overgrowth of forebrain tissue (Fig. 4e) both in the ventral forebrain and cerebral cortex, characterized by heterotopic foci of proliferation and aberrant cortical foldings (Fig. 4f).

Together, the data indicate that ephrin/Eph signalling triggers pro-apoptotic pathways in early cortical progenitors, thereby controlling their number, and hence the final size of the forebrain. Although the prominent role of programmed cell death of post-mitotic neurons in brain development has been well established, the physiological meaning of apoptosis in neural progenitors has remained less clear¹¹. Its influence on brain size was revealed by the analysis of mice in which apoptosis effectors and regulators have been knocked out: caspase-3-9 and Apaf1 mutants all display a severe exencephalic overgrowth of brain structures and reduction in neural progenitor apoptosis^{11,12}, strikingly reminiscent of our observations in some EphA7^{-/-} mutants (Fig. 4e, f). Although the low penetrance of this extreme phenotype did not allow us to establish a strict correlation with abnormal apoptosis rates, it probably results from a more severe dysregulation of cell death at earlier stages of forebrain development, as previously described for caspase mutants. A recent study identified caspases-3-9 among genes that underwent positive selection during primate evolution¹⁶, suggesting that regulation of cell death might be one of the mechanisms underlying the expansion of the human cerebral cortex. Region-specific apoptosis of neural progenitors has also recently been proposed as a mechanism to restrict in time and space the size of distinct neural stem cell lineages in the *Drosophila* central nervous system¹⁷. In this context, future work will be needed to determine which ephrin ligand(s) interact physiologically with EphA7 during early cortical neurogenesis. Ephrin-A5 is present in the ventral telencephalon (see Fig. 1 and Supplementary Fig. 3), suggesting that tangentially migrating ventral neurons could display ephrin-A5 in the developing cortex. Ephrin-A3, expressed early in the cortical plate, and ephrin-A2, expressed in the ventricular zone (Supplementary Fig. 3), could also act synergistically as cortical EphA7-ligands to provide spatial control of neural progenitor apoptosis.

Ephrins have previously been proposed to control cell survival *in vitro*^{18,19}, but with no physiological implication of ephrins in programmed cell death. Recent data have implicated apoptosis effectors and regulators in the control of axon or cell guidance^{20,21}. Conversely, other cell and axon guidance factors have been shown to be involved in the control of cell death during development and in oncogenesis²²⁻²⁶. In most cases, apoptosis was shown to result from lack of activation of the receptors by their ligand, leading to the concept of dependence receptors. However, in the case of ephrins, apoptosis seems to result from direct stimulation by the ligand, similarly to Fas-ligand or neurotrophin-induced cell death^{27,28}.

Ephrin/Eph signalling was recently reported to mediate an anti-proliferative effect on adult neural progenitors, and thereby to influence neurogenesis in the adult olfactory bulb²⁹. However, our study demonstrates a direct pro-apoptotic effect of ephrins on embryonic cortical progenitors, with no influence on their proliferation. This further illustrates that ephrins, like neurotrophins, have evolved as pleiotropic factors that can control very different functions depending on the cellular context. Although the mechanisms allowing ephrin signalling to control guidance, apoptosis or proliferation in a cell-specific manner remain unknown, the identification of ephrin/Eph receptor genes as positive regulators of apoptosis during forebrain neurogenesis uncovers a novel signalling pathway that could potentially be involved in other aspects of developmental or stem cell biology, and in oncogenesis.

METHODS

Transgenic mice. Generation and genotyping of Emx1^{IRESc} (Emx1-Cre)

knock-in and EphA7 knockout mice were described previously^{9,13}. For the generation of TGA7A5 mice, a BAC homologous recombination method was used^{7,8} (see Supplementary Information for details). A BAC clone of 200 kilobases (kb), centred on the *Epha7* gene, was targeted with a conditional cassette containing an eGFP expression cassette flanked by loxP sites, followed by Flag-tagged human ephrin-A5 complementary DNA and a kanamycin-resistance cassette (Fig. 1b). Homologous recombinant BACs, in which the first exon was replaced by the conditional cassette, were obtained using established methods⁷. TGA7A5 mice were generated using microinjection of the recombinant BAC in C57BL/6 × CBA F₁ zygotes and were screened using Southern blotting and polymerase chain reaction (PCR).

In situ RNA hybridization and immunohistochemistry. *In situ* RNA hybridization using digoxigenin-labelled RNA probes on fixed brain cryosections was performed as described³⁰. Immunodetection was performed using standard methods (antibodies used are described in the Supplementary Methods). All stainings were performed on at least three embryos from each genotype tested.

Quantitative studies of cell proliferation, cell cycle exit and cortical size. For BrdU labelling, timed-pregnant female mice were injected intraperitoneally with a single pulse (50 mg kg⁻¹ body weight) of BrdU, killed after 1 h or 24 h, and embryos fixed by perfusion with 4% paraformaldehyde. Simultaneous staining of Ki67 and BrdU, and quantification of labelling index and quitting fraction were performed according to the methods in ref. 10. For each brain analysed, at least 400 cells, distributed across four sections, were counted. To quantify cortical size, cortical area was measured on two sets of six sections for each animal, corresponding to defined fiduciary marks (rostral pole of the thalamus, caudal pole of the internal capsule). Statistical analyses were performed using Student's *t* test.

Culture of dissociated cortical progenitors. Cortex was dissected from E12.5 embryos and dissociated mechanically in L15 medium (Gibco) containing 30 mM glucose. The suspension was centrifuged at 2,000g for 4 min at 20 °C and the pellet resuspended in culture medium (70% basal Eagle medium, 25% Hanks' balanced salt solution, 25 mM glucose, 1 mM glutamine, 50 U ml⁻¹ penicillin, 50 µg ml⁻¹ streptomycin, 5% horse serum). Dissociated cortical cells were plated at a density of 6–8 × 10⁵ cells per well in 6-well plates pre-coated with poly-D-lysine (33 µg ml⁻¹) and laminin (3 µg ml⁻¹). Cultures were maintained for 24 h in a humidified atmosphere at 5% CO₂, 37 °C, then stimulated with 100 nM preclustered ephrin-A5-Fc or control Fc fragment, with or without 40 µM z-VAD-fmk (R&D Systems). After treatment, cells were either fixed in 4% paraformaldehyde in PBS for 15 min at room temperature (for TUNEL assays) or trypsinized (for caspase-3 activity assays).

TUNEL detection and caspase activity assays. TUNEL assays were performed using a commercially available kit and following to manufacturer's instructions (TMR red, Roche Diagnostics). The number of apoptotic cells (detected by TUNEL assay and activated caspase-3 immunoreactivity) was counted throughout the cerebral cortex in EphA7^{-/-} and EphA7^{+/+} littermates, and normalized to the number of brain sections analysed. For *in vitro* assays, the percentage of apoptotic cells was determined by counting the number of TUNEL-positive cells in 1,000 cells (visualized by Hoechst staining) for each duplicate from each independent experiment. Enzymatic assays for caspase-3 activity were performed using a commercially available kit according to manufacturer's instructions (ApoAlert, BD Biosciences). Statistical analyses were performed using Student's *t* test.

Received 18 February; accepted 20 April 2005

Published online 15 May 2005.

1. Caviness, V. S. Jr, Takahashi, T. & Nowakowski, R. S. Numbers, time and neocortical neurogenesis: a general developmental and evolutionary model. *Trends Neurosci.* **18**, 379–383 (1995).
2. Mochida, G. H. & Walsh, C. A. Molecular genetics of human microcephaly. *Curr. Opin. Neurol.* **14**, 151–156 (2001).
3. Flanagan, J. G. & Vanderhaeghen, P. The ephrins and Eph receptors in neural development. *Annu. Rev. Neurosci.* **21**, 309–345 (1998).
4. Klein, R. Eph/ephrin signaling in morphogenesis, neural development and plasticity. *Curr. Opin. Cell Biol.* **16**, 580–589 (2004).
5. Poliakov, A., Cotrina, M. & Wilkinson, D. G. Diverse roles of eph receptors and ephrins in the regulation of cell migration and tissue assembly. *Dev. Cell* **7**, 465–480 (2004).
6. Yun, M. E., Johnson, R. R., Antic, A. & Donoghue, M. J. EphA family gene expression in the developing mouse neocortex: regional patterns reveal intrinsic programs and extrinsic influence. *J. Comp. Neurol.* **456**, 203–216 (2003).
7. Chaveroche, M. K., Ghigo, J. M. & d'Enfert, C. A rapid method for efficient gene replacement in the filamentous fungus *Aspergillus nidulans*. *Nucleic Acids Res.* **28**, E97 (2000).

8. Zhang, Y., Muylers, J. P., Testa, G. & Stewart, A. F. DNA cloning by homologous recombination in *Escherichia coli*. *Nature Biotechnol.* **18**, 1314–1317 (2000).
9. Gorski, J. A. *et al.* Cortical excitatory neurons and glia, but not GABAergic neurons, are produced in the Emx1-expressing lineage. *J. Neurosci.* **22**, 6309–6314 (2002).
10. Chenn, A. & Walsh, C. A. Regulation of cerebral cortical size by control of cell cycle exit in neural precursors. *Science* **297**, 365–369 (2002).
11. Kuan, C. Y., Roth, K. A., Flavell, R. A. & Rakic, P. Mechanisms of programmed cell death in the developing brain. *Trends Neurosci.* **23**, 291–297 (2000).
12. Kuida, K. *et al.* Decreased apoptosis in the brain and premature lethality in CPP32-deficient mice. *Nature* **384**, 368–372 (1996).
13. Dufour, A. *et al.* Area specificity and topography of thalamocortical projections are controlled by ephrin/Eph genes. *Neuron* **39**, 453–465 (2003).
14. Wilkie, A. L., Jordan, S. A., Sharpe, J. A., Price, D. J. & Jackson, I. J. Widespread tangential dispersion and extensive cell death during early neurogenesis in the mouse neocortex. *Dev. Biol.* **267**, 109–118 (2004).
15. Thomaidou, D., Mione, M. C., Cavanagh, J. F. & Parnavelas, J. G. Apoptosis and its relation to the cell cycle in the developing cerebral cortex. *J. Neurosci.* **17**, 1075–1085 (1997).
16. Dorus, S. *et al.* Accelerated evolution of nervous system genes in the origin of *Homo sapiens*. *Cell* **119**, 1027–1040 (2004).
17. Bello, B. C., Hirth, F. & Gould, A. P. A pulse of the *Drosophila* Hox protein Abdominal-A schedules the end of neural proliferation via neuroblast apoptosis. *Neuron* **37**, 209–219 (2003).
18. Dohn, M., Jiang, J. & Chen, X. Receptor tyrosine kinase EphA2 is regulated by p53-family proteins and induces apoptosis. *Oncogene* **20**, 6503–6515 (2001).
19. Yue, Y. *et al.* Selective inhibition of spinal cord neurite outgrowth and cell survival by the Eph family ligand ephrin-A5. *J. Neurosci.* **19**, 10026–10035 (1999).
20. Campbell, D. S. & Holt, C. E. Apoptotic pathway and MAPKs differentially regulate chemotropic responses of retinal growth cones. *Neuron* **37**, 939–952 (2003).
21. Geisbrecht, E. R. & Montell, D. J. A role for *Drosophila* IAP1-mediated caspase inhibition in Rac-dependent cell migration. *Cell* **118**, 111–125 (2004).
22. Bagnard, D. *et al.* Semaphorin 3A-vascular endothelial growth factor-165 balance mediates migration and apoptosis of neural progenitor cells by the recruitment of shared receptor. *J. Neurosci.* **21**, 3332–3341 (2001).
23. Thiebault, K. *et al.* The netrin-1 receptors UNC5H are putative tumor suppressors controlling cell death commitment. *Proc. Natl Acad. Sci. USA* **100**, 4173–4178 (2003).
24. Williams, M. E., Strickland, P., Watanabe, K. & Hinck, L. UNC5H1 induces apoptosis via its juxtamembrane region through an interaction with NRAGE. *J. Biol. Chem.* **278**, 17483–17490 (2003).
25. Matsunaga, E. *et al.* RGM and its receptor neogenin regulate neuronal survival. *Nature Cell Biol.* **6**, 749–755 (2004).
26. Mazelin, L. *et al.* Netrin-1 controls colorectal tumorigenesis by regulating apoptosis. *Nature* **431**, 80–84 (2004).
27. Raoul, C., Pettmann, B. & Henderson, C. E. Active killing of neurons during development and following stress: a role for p75(NTR) and Fas? *Curr. Opin. Neurobiol.* **10**, 111–117 (2000).
28. Lee, R., Kerami, P., Teng, K. K. & Hempstead, B. L. Regulation of cell survival by secreted proneurotrophins. *Science* **294**, 1945–1948 (2001).
29. Holmberg, J. *et al.* Ephrin-A2 reverse signaling negatively regulates neural progenitor proliferation and neurogenesis. *Genes Dev.* **19**, 462–471 (2005).
30. Vanderhaeghen, P. *et al.* A mapping label required for normal scale of body representation in the cortex. *Nature Neurosci.* **3**, 358–365 (2000).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank G. Vassart for continuous support and interest, and members of IRIBHM for help and advice, in particular D. Communi, N. Gaspard, M.-A. Lambot, S. Pietri, J. Seibt and F. Pauwels. We also thank the Development Studies Hybridoma Bank; T. Ciossek for sharing the EphA7 mutant mice; C. d'Enfert and J.-M. Ghigo for the gift of the pKOBEGA plasmid; A. de Kerchove for discussions about BAC recombineering; A. Deplano for help with pulse-field electrophoresis; S. Tajbakhsh for the gift of the floxed eGFP cassette; F. Stewart for the gift of the frt-Neo cassette and the 294-Flp bacteria; A. Goffinet for the gift of mouse reelin plasmid; R. Hevner for the gift of Tbr1 antibody; and F. Polleux for critically reading the manuscript. This work was funded by grants from the Belgian Funds for Scientific Research (FNRS and FRSM), the Belgian Queen Elizabeth Medical Foundation, the Belgian Interuniversity Attraction Poles Programme (to P.V.), and grants from the American Cancer Society and National Institutes of Health (to K.R.J.). P.V. and C.L. are Research Associates, and L.P. is a Research Fellow, of the FNRS. A.D. and V.D. were supported by the Fonds pour la Recherche en Industrie et Agriculture (FRIA).

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to P.V. (pvdhaegh@ulb.ac.be).

Auxin inhibits endocytosis and promotes its own efflux from cells

Tomasz Paciorek¹, Eva Zažímalová², Nadia Ruthardt³, Jan Petrášek², York-Dieter Stierhof¹, Jürgen Kleine-Vehn¹, David A. Morris^{2,4}, Neil Emans³, Gerd Jürgens¹, Niko Geldner¹ & Jiří Friml¹

One of the mechanisms by which signalling molecules regulate cellular behaviour is modulating subcellular protein translocation. This mode of regulation is often based on specialized vesicle trafficking, termed constitutive cycling, which consists of repeated internalization and recycling of proteins to and from the plasma membrane¹. No such mechanism of hormone action has been shown in plants although several proteins, including the PIN auxin efflux facilitators, exhibit constitutive cycling^{2,3}. Here we show that a major regulator of plant development, auxin, inhibits endocytosis. This effect is specific to biologically active auxins and requires activity of the Calossin-like protein BIG. By inhibiting the internalization step of PIN constitutive cycling, auxin increases levels of PINs at the plasma membrane. Concomitantly, auxin promotes its own efflux from cells by a vesicle-trafficking-dependent mechanism. Furthermore, asymmetric auxin translocation during gravitropism is correlated with decreased PIN internalization. Our data imply a previously undescribed mode of plant hormone action: by modulating PIN protein trafficking, auxin regulates PIN abundance and activity at the cell surface, providing a mechanism for the feedback regulation of auxin transport.

The local, asymmetric distribution of the plant growth regulator auxin mediates a variety of developmental processes such as axis formation, organ initiation and positioning, directional growth (tropisms) and meristem activity^{4–6}. Biochemical, genetic and molecular data have confirmed that auxin promotes SCF^{TIR1}-mediated ubiquitination and degradation of the auxin/indole-3-acetic acid (AUX/IAA) repressors, thus releasing auxin response factor (ARF) transcriptional regulators from inhibition⁷. In this manner the expression of different sets of genes is activated, thereby eliciting different cellular and, consequently, developmental responses. An important additional level of regulation upstream of cellular auxin signalling is a specific transport system dependent on polarly localized PIN auxin efflux regulators⁸. This dynamic auxin distribution network mediates directional (polar) auxin flow between cells, which contributes to the formation and maintenance of asymmetric auxin distribution.

PIN proteins rapidly and constitutively cycle between the plasma membrane (PM) and endosomes². In *Arabidopsis* this cycling involves a brefeldin A (BFA)-sensitive regulator of vesicle budding, the guanine exchange factor for adenosine-ribosylation-factor-type small GTPases (ARF GEF) known as GNOM³. BFA inhibits trafficking from endosomes to the PM and causes endosomes to aggregate into 'BFA compartments'², which become surrounded by Golgi stacks³. In contrast, trafficking from the PM to the endosomes seems to be insensitive to BFA. These differential effects of BFA on endocytosis and exocytosis lead to the internalization and

accumulation of constitutively cycling proteins in the BFA compartments. Thus, in *Arabidopsis* BFA can serve as a tool for revealing subcellular protein movement between endosomes and the PM^{2,3}.

Several plant PM markers were rapidly and reversibly internalized in response to BFA. These included PIN auxin efflux regulators such as PIN1 (ref. 2) (Fig. 1a, b), PIN2 (ref. 3) (Supplementary Fig. S1a), PIN3 (ref. 4) and PIN4 (Supplementary Fig. S1d), the PM water channel PIP2 (Fig. 1d) and maize cell-wall pectins⁹ (Supplementary Figure S1h). Plasma membrane H⁺-ATPase (PM-ATPase) was preferentially internalized in rapidly elongating epidermal cells (Fig. 1c), indicating the possible existence of both dynamic and more static populations of the protein. BFA-induced reversible internalization also occurred when transcription, protein synthesis² or protein degradation were inhibited (Supplementary Fig. S2), confirming that proteins accumulating in BFA compartments were not synthesized *de novo* but originated from the PM. Co-localization of cycling proteins (Supplementary Fig. S1a–c), the endocytic tracer FM4-64 (Supplementary Fig. S1e–g) and endosomal markers (Fig. 2b) revealed that the proteins studied here recycle through the same endomembrane compartments. These data show that some plant PM proteins are retained at the PM, but many exhibit constitutive cycling between the PM and endosomes.

In animals, constitutive cycling is an entry point for multiple regulation, including by signalling molecules¹. Through this mechanism, hormones such as insulin or vasopressin can control the relative rates of endocytosis and exocytosis and thereby regulate the concentrations, and thus the activity of surface-localized proteins, including ion and water channels, transporters and receptors¹. In plants, even high concentrations of several phytohormones including abscisic acid, brassinosteroids, cytokinins, ethylene and gibberellins had no detectable effect on various trafficking processes, including BFA-induced internalization (Fig. 1h, Supplementary Fig. S3). Interestingly, however, auxins such as the naturally occurring IAA and its synthetic analogues naphthalene-1-acetic acid (NAA) and 2,4-dichlorophenoxyacetic acid (2,4-D) efficiently inhibited the internalization of PIN1, PIN2, PIN4, PM-ATPase, PIP2:GFP and maize cell wall pectins in response to BFA (Fig. 1e–g, Supplementary Fig. S1i–p). Concentrations as low as 5 μ M NAA and 2,4-D were enough to elicit near-maximal effects (Fig. 1e, f, Supplementary Fig. S4) indicating that these compounds might be sufficiently active even at lower concentrations. IAA was chemically unstable under our experimental conditions (verified by high-performance liquid chromatography and mass spectroscopy; Supplementary Fig. S5), but when an antioxidant (butylated hydroxytoluene; BHT¹⁰) was included in the medium, IAA was also effective at concentrations as low as 5 μ M (Fig. 1g, Supplementary Fig. S5). In contrast, the physiologically inactive structural isomer of

¹Zentrum für Molekularbiologie der Pflanzen, Universität Tübingen, 72076 Tübingen, Germany. ²Institute of Experimental Botany, ASCR, Rozvojová 135, 165 02 Praha 6, Czech Republic. ³Cellome Research Group, RWTH Aachen University, Biology VII, Worringerweg 1, 52074 Aachen, Germany. ⁴School of Biological Sciences, University of Southampton, Bassett Crescent East, Southampton SO16 7PX, UK.

NAA, naphthalene-2-acetic acid (2-NAA), had no effect even at a tenfold higher concentration (Fig. 1i). However, even at maximal effective concentrations, auxins did not completely block BFA-induced internalization. This is inferred from the observation that some accumulation of constitutively cycling PM proteins inside cells

could be observed when the treatment with BFA in the presence of auxins was prolonged (Supplementary Fig. S6).

Induced increases in intracellular concentrations of endogenous auxin also had an inhibitory effect on the internalization of constitutively cycling PM proteins. *35S::TMS2* plants overexpress the amidohydrolase that converts biologically inactive auxin amides into active auxins¹¹. When treated with NAA amide or IAA amide, *35S::TMS2* (Fig. 1j) but not wild-type plants (data not shown) showed inhibition of BFA-induced PIN1 internalization. Furthermore, *Arabidopsis* mutants with increased concentrations of endogenous auxin such as *superroot1* (*sur1*)¹² (data not shown) or *yucca*¹³ (Fig. 1k) showed decreased PIN1 internalization after treatment with BFA in comparison with wild-type plants. Protophloem cells of root have been shown to have higher concentrations of auxin than surrounding tissues¹⁴. In wild-type plants grown under normal conditions, BFA-induced internalization was clearly inhibited in these cells (Fig. 1l). Together, these results show that physiological concentrations of exogenously applied and/or endogenously produced auxins downregulate the BFA-induced internalization of constitutively cycling PM proteins.

Next we examined the subcellular site(s) at which auxin acts on the BFA-induced internalization of PM proteins. An important control was the confirmation that auxins did not influence the uptake of BFA into *Arabidopsis* roots (Supplementary Fig. S7). To assess possible effects of auxin on different trafficking processes, we used established markers for various endomembrane compartments. There were no apparent effects of auxin on the distribution of ER (Sec12 (ref. 3)), presumptive trans-Golgi network (TLG2a (ref. 3)), Golgi apparatus (γ -COP (ref. 3)) or endosomal (ARF1 (ref. 15)) markers (Supplementary Fig. S8). Auxins also did not interfere with the BFA-induced formation of endosomal BFA compartments or with the aggregation of Golgi stacks at their periphery (Fig. 2a–f, Supplementary Fig. S9). Double labelling revealed that, in the same cells, auxin treatment did not affect the formation of Golgi-stack-encircled BFA compartments but prevented the internalization of PM proteins and therefore their accumulation in such compartments (Fig. 2c, f). The observation that BFA compartments (revealed as an aggregation of endocytic vesicles) still formed after treatment with auxin was confirmed by an examination of ultrastructure by electron microscopy (Fig. 2k–m). These data indicate that auxins might interfere with the endocytic step of constitutive cycling without visibly affecting other subcellular trafficking processes.

To assess directly the effect of auxin on endocytosis, we used the fluorescent dye FM4-64, an established endocytic tracer³. In *Arabidopsis* roots, even low concentrations of exogenously applied auxins clearly decreased the detectable uptake of FM4-64 (Fig. 2g, h), showing the inhibition of endocytosis. Furthermore, in *yucca* roots, which contain higher concentrations of endogenous IAA¹³, the uptake of FM4-64 was also inhibited (Fig. 2i). To confirm and quantify the auxin effect on endocytosis, we measured the uptake of another endocytic tracer, FM1-43 (ref. 16), into suspension-grown tobacco BY-2 cells. Both 2,4-D and NAA inhibited FM1-43 uptake in a concentration-dependent manner (Fig. 2j). NAA was less effective, which is consistent with its decreased retention in tobacco cells¹⁷. Auxins also completely abolished the BFA-induced increase in FM1-43 internalization that results from the inhibition of membrane recycling back to the cell surface¹⁶. These experiments confirmed that the endocytic step of the cycling of PM proteins is the target for auxin action.

For a molecular characterization of the pathway by which auxin inhibits endocytosis, we performed a genetic screen to find mutants altered in the auxin effect on endocytosis. One group of mutants that showed resistance to the auxin effect on endocytosis was allelic to *transport inhibitor response3* (*tir3*). *tir3* was originally isolated in a screen for resistance to auxin transport inhibitors¹⁸ and other alleles have also been identified by their involvement in light signal

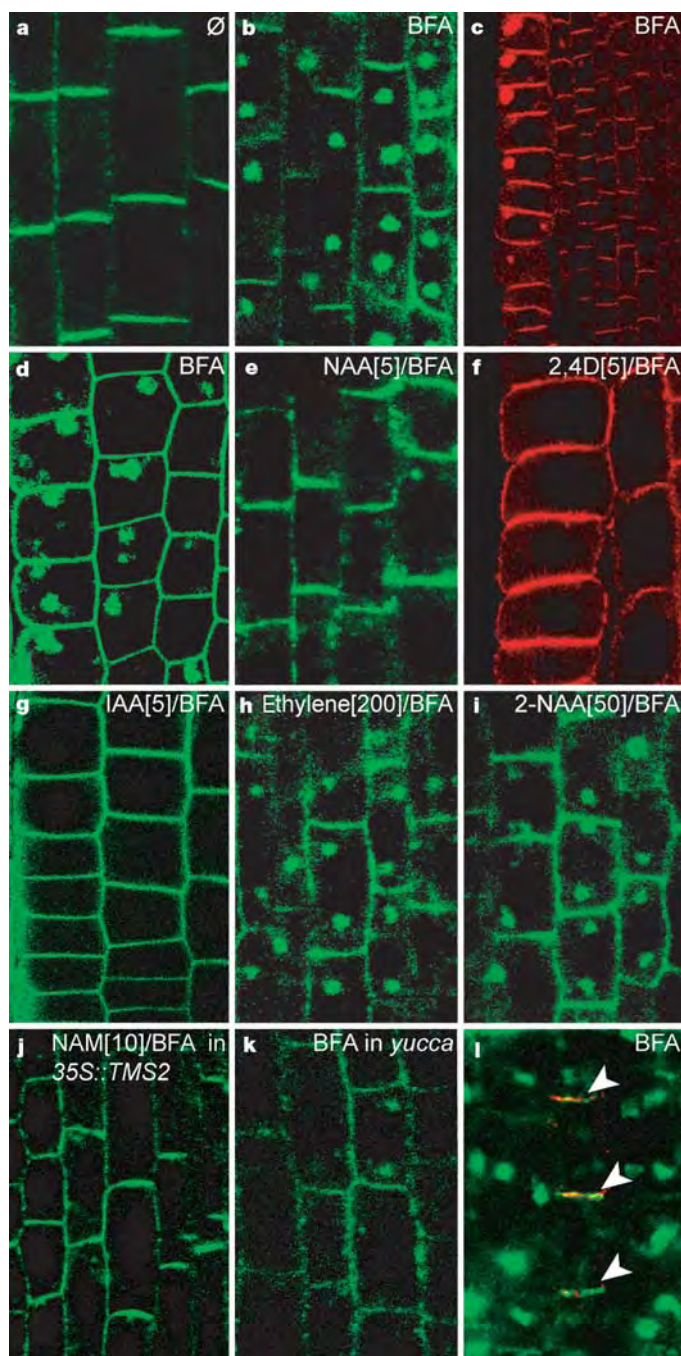


Figure 1 | Auxins inhibit internalization of constitutively cycling proteins. **a–d**, BFA-induced internalization of PM proteins: Polar, PM localization of PIN1 in untreated ($\emptyset$) roots (**a**); PIN1 (**b**) and PIP2:GFP (**d**) internalization. BFA-sensitive (epidermis, cortex) and BFA-insensitive (stele) pools of PM-ATPase (**c**). **e–g**, Auxins (NAA, 2,4-D and IAA) inhibit BFA-induced internalization of PIN1 (**e**), PM-ATPase (**f**) and PIP2:GFP (**g**). **h, i**, Other hormones including ethylene (aminocyclopropane carboxylic acid (**h**)) or the inactive NAA analogue 2-NAA (**i**) are ineffective. **j**, NAA amide (NAM) in *35S::TMS2* plants inhibits BFA-induced internalization. **k**, *yucca* mutants show decreased PIN1 internalization. **l**, Protophloem cells marked by AUX1 (red) show less BFA-induced PIN1 (green) internalization than surrounding cells. Numbers in square brackets are concentrations in μ M.

transduction (dark overexpression of *CAB1* (*doc1*) and attenuated shade avoidance (*asa1*)) or cytokinin response (*umbrella* (*umb1*))^{19,20}. The corresponding gene, since renamed *BIG*, encodes a member of the Calossin/Pushover family present in other multicellular organisms¹⁹. These proteins might be involved in subcellular vesicle trafficking, because they affect the subcellular localization of PIN1 after treatment with auxin efflux inhibitors¹⁹. We found that in both the *tir3* and *doc1* alleles of *big*, endocytosis of PM proteins was inhibited by auxins less than in the wild type. Auxin concentrations that would normally lead to the visible inhibition of BFA-induced internalization of the PIN1 protein (Fig. 3a–c) were ineffective in *big* mutants (Fig. 3e–g). Only higher auxin concentrations were able to

inhibit the BFA-induced internalization (Fig. 3d, h). Other tested PM proteins such as PIN2 (data not shown) and PM-ATPase (Fig. 3i) behaved in a similar way as PIN1. Furthermore, *big* mutants also showed resistance to endogenously increased auxin concentrations in combination with *yucca* mutants, as demonstrated by a pronounced inhibition of BFA-induced internalization in *yucca* (see Fig. 1k), but not in *yucca big* mutant roots (Fig. 3j). In addition, the inhibitory effect of auxin on the internalization of the endocytic tracer FM4-64 (see Fig. 2g–i) was less pronounced in *big* mutant roots (Fig. 3k). These data indicate that *BIG* is required for the auxin-mediated inhibition of endocytosis and thereby identify a molecular component of this specific pathway of auxin action.

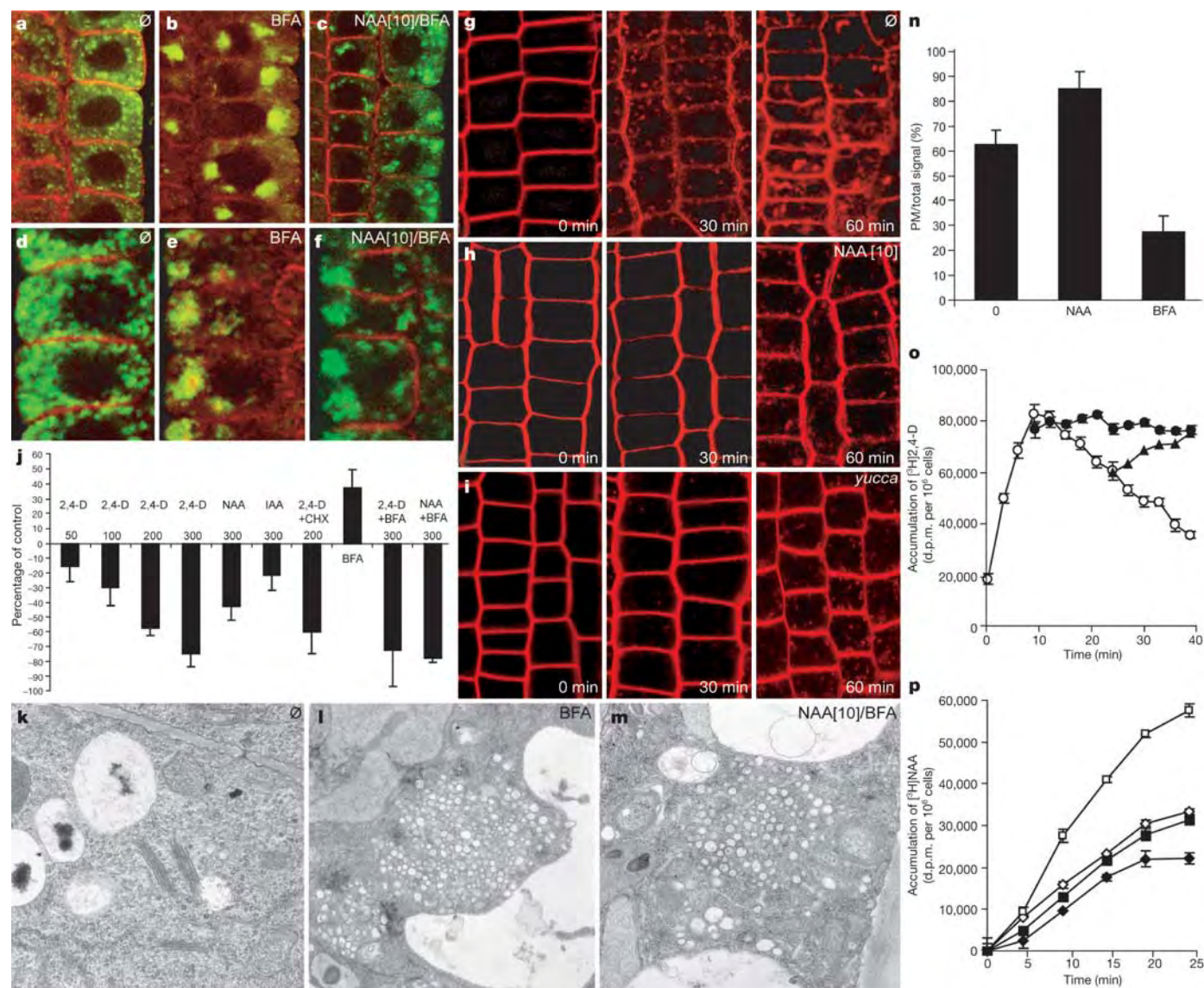


Figure 2 | Auxin inhibits endocytosis, increases the amount of PIN2 protein at the plasma membrane and stimulates its own efflux from tobacco cells. **a–c**, PM-ATPase (red) is internalized in response to BFA into ARF1-containing (green), endosomal BFA compartments (**b**). After treatment with NAA, PM-ATPase does not internalize, but ARF1-containing endosomes aggregate (**c**). **d–f**, γ -COP-labelled Golgi apparatus (green) aggregates around PM-ATPase-containing BFA compartments (red) (**e**). NAA prevents PM-ATPase internalization but not γ -COP aggregation (**f**). **g–i**, Uptake of endocytic tracer FM4-64 is slowed down in NAA-treated roots (**h**) and in *yucca* roots (**i**). **j**, 2,4-D, NAA and IAA inhibit uptake of FM1-43 into BY-2 cells in a concentration-dependent manner. Concentrations shown are in μ M. CHX, cycloheximide. **k–m**, Ultrastructural examination of BFA compartments did not reveal any differences in ultrastructure between

BFA (**l**) and NAA/BFA (**m**) treatments. **n**, The ratio of PM-located to internalized PIN2 signals (in %) increases after NAA and decreases after BFA treatments. **o**, [3 H]2,4-D decreases its own accumulation in VBI-0 cells over time (open circles). Concomitant BFA treatment (applied 6 min after addition of [3 H]2,4-D) prevents this effect (filled circles), and NAA (applied 21 min after addition of [3 H]2,4-D) competitively inhibits [3 H]2,4-D export (triangles). **p**, Pretreatment with 2,4-D for 20 min (filled diamonds) decreases [3 H]NAA accumulation in the BY-2 cells. BFA increases the accumulation (open squares), but this effect is prevented by pretreatment with 2,4-D (filled squares). Open diamonds, no addition. **a, d, g, k**, Untreated controls. Numbers in square brackets are concentrations in μ M. Error bars show s.d.; $\emptyset$ relates to untreated controls.

If auxins inhibit the endocytic step of constitutive cycling, the protein pool at the PM should increase after treatment with auxin. To test this prediction we directly measured the ratio of cell surface to internalized PIN2 protein by using quantitative confocal microscopy. In the control (no auxin treatment), the PM pool composed approximately 62% of the total PIN2 protein. Treatment with auxins (NAA) increased the cell-surface signal significantly (85%, $P < 0.001$), showing higher levels of PIN2 at the PM (Fig. 2n). By contrast, treatment with BFA (inhibiting the exocytic step of the cycling) had the opposite effect and caused a massive accumulation of PIN2 inside cells, decreasing the cell-surface signal (28%, $P < 0.001$). Taken together, these data strongly indicate that auxins might inhibit the endocytic step of constitutively cycling PIN proteins, thus increasing their levels at the PM.

An auxin-dependent increase in the amount of cell-surface-localized PIN auxin efflux regulators would afford a mechanism by which auxin can control its own transport. A classical model, which attempts to explain multiple self-organizing auxin effects—the so-called canalization hypothesis—proposed feedback regulation between auxin signalling and intercellular auxin transport²¹. Earlier physiological experiments did indeed imply that auxin is required to maintain its own polar transport²². However, a direct effect of auxin on its own efflux has not been shown. The effects of auxin on its own transport were quantitatively assessed in suspension-grown tobacco cell lines VBI-0 and BY-2, which are well-established systems for studies of auxin transport *in vivo*^{23,24}. The lines differ in their abilities

to export different auxins, and the most important feature related to this study is that 2,4-D is a much better substrate for an auxin efflux machinery in VBI-0 cells than in BY-2 cells (Supplementary Fig. S10a–c). Generally, the net accumulation of radioactively labelled auxins in cells gives a measure of the relative rates of their uptake and efflux. In constant carrier-driven (and thus saturable) transport across the PM, the auxin accumulation would be expected to exhibit saturation kinetics until the uptake and efflux rates reached equilibrium. When VBI-0 cells were incubated with [³H]2,4-D, after an initial increase in its internal concentration, the accumulation of [³H]2,4-D decreased steadily with time instead of reaching a stable equilibrium (Fig. 2o). This effect was completely reversed by both a known inhibitor of auxin efflux (1-naphthylphthalamic acid (NPA); Supplementary Fig. S10d) and a high-affinity substrate for efflux carrier(s) in tobacco cells (NAA^{17,23}; Fig. 2o). This shows that the observed decrease in 2,4-D accumulation results from an increased capacity in its carrier-driven efflux. This decrease in accumulation

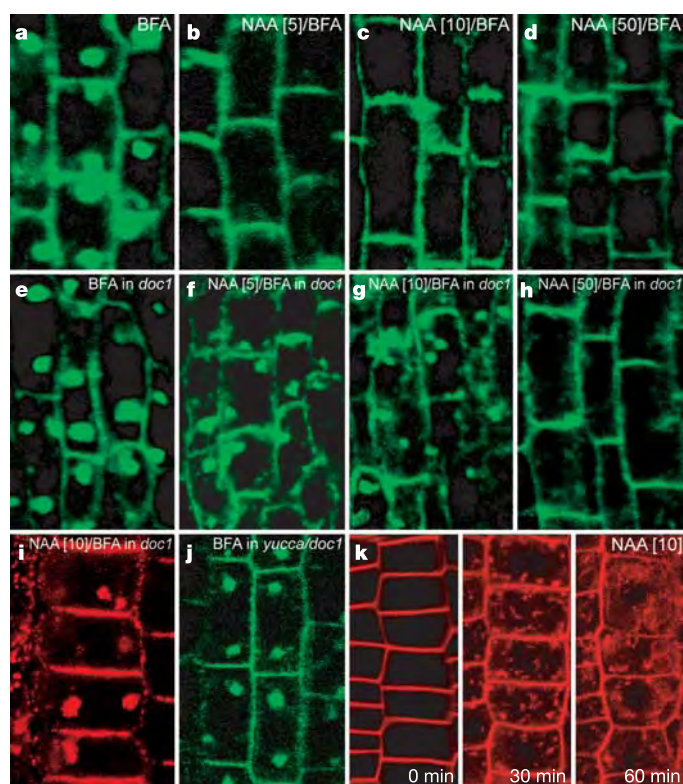


Figure 3 | BIG protein is required for the auxin-dependent inhibition of endocytosis. a–h, *doc1* mutants are partly auxin-resistant: BFA causes PIN1 internalization in control (a) and *doc1* roots (e). NAA blocks PIN1 internalization in controls (b, c) but not in *doc1* (f, g). Higher NAA concentrations are effective in both control (d) and *doc1* (h). i, In *doc1*, BFA-induced internalization of PM-ATPase is resistant to auxin (compare with Fig. 1f). j, Increase in auxin concentrations in *yucca* does not inhibit PIN1 internalization in the *yucca doc1* double mutant (compare with Fig. 1k). k, NAA does not inhibit the uptake of FM4-64 in the *doc1* mutant (compare with Fig. 2g, h). Numbers in square brackets are concentrations in μ M.

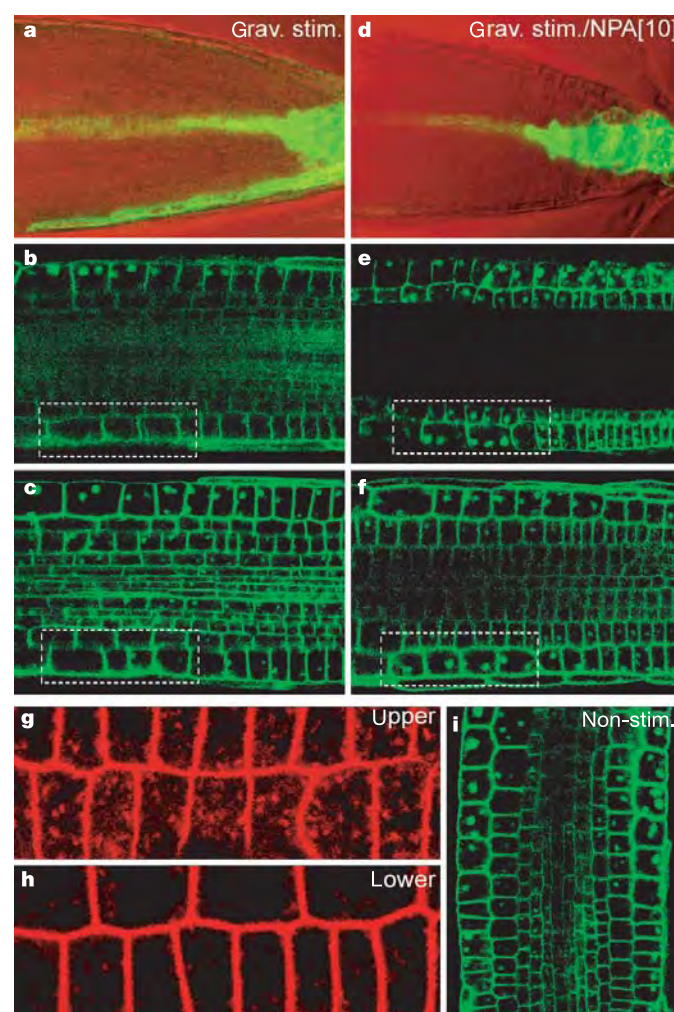


Figure 4 | Correlation between auxin translocation and the rate of PIN2 internalization in course of root gravitropism. a–c, After stimulation by gravity, auxin translocation detected by DR5rev::GFP activity (a) at the lower side of roots is correlated with the inhibition of BFA-induced internalization of PIN2::GFP (b) and PIP2::GFP (c). d–f, When auxin translocation is prevented by NPA, no asymmetry in DR5rev::GFP expression (d) or BFA-induced internalization of PIN2::GFP (e) and PIP2::GFP (f) occurs. g, h, After stimulation by gravity, FM4-64 uptake is lower in layers on the underside of the root (h) than in those on the upper side (g). i, Without stimulation by gravity, no asymmetry in the BFA-induced internalization of PIP2::GFP was detected. Numbers in square brackets are concentrations in μ M.

(that is, an increase in efflux) was abolished by treatment with BFA (Fig. 2o), indicating that the stimulation of 2,4-D efflux was dependent on BFA-sensitive recycling of auxin efflux carriers back to PM. Next we investigated whether the pretreatment with auxin could directly affect the activity of auxin efflux. In this case, the accumulation of auxin had to be measured by using an auxin that is taken up by passive diffusion (to exclude the possible interference by the activity of auxin influx carriers) and, simultaneously, one that is a good substrate for efflux carrier(s) in tobacco cells; NAA fulfilled these criteria in both VBI-0 and BY-2 cells. In contrast, for the pretreatment, an auxin that is a weak substrate for an auxin efflux carrier had to be used to prevent competitive inhibition of NAA efflux. Because in VBI-0 cells, but not in BY-2 cells, 2,4-D is a good substrate for an efflux carrier (Supplementary Fig. S10a–c), we used BY-2 cells for this experiment. Pretreatment for 20 min with 5 μ M 2,4-D significantly decreased the accumulation of [3 H]NAA (increased its efflux) and, moreover, abolished the inhibitory effect of BFA on auxin efflux (Fig. 2p). Together, these data show that in tobacco cells *in vivo*, auxin stimulates its own efflux by a vesicle-trafficking-dependent mechanism. All the results are consistent with the notion that auxin inhibits the internalization of constitutively cycling auxin efflux catalysts, thus increasing their incidence at the cell surface and thereby stimulating its own efflux.

We next explored *in planta* the link between the effects of auxin on endocytosis and auxin transport. Root gravitropism—the directional growth of roots along a gravity vector—is a well-characterized physiological process that involves the rapid establishment of auxin flow along the lower side of a horizontal root after gravity perception^{4,8}. This process represents an ideal system in which to examine auxin translocation and endocytic recycling in parallel. After stimulation by gravity, asymmetric auxin flow resulted in auxin accumulation at the lower side of the root, as detected by the auxin response reporter DR5rev::GFP (Fig. 4a). Significantly, a similar spatial asymmetry within the root was observed for the BFA-induced internalization of PIN2:GFP and PIP2:GFP (Fig. 4b, c), indicating that their endocytic recycling might have been inhibited at the lower side of the gravistimulated root. Reduced uptake of the endocytic tracer FM4-64 was also observed at the lower side, in comparison with the upper side, of the root (Fig. 4g, h). In contrast, no asymmetry in auxin accumulation (as detected by DR5rev::GFP; Fig. 4d) and no inhibition of PM protein internalization in response to BFA occurred in non-stimulated roots (Fig. 4i) or after treatment with NPA (Fig. 4e, f), which under these conditions inhibits auxin flow but not protein cycling². Thus, asymmetric auxin translocation is closely correlated spatially with the inhibition of endocytosis during the root gravitropic response. This finding strengthens the likelihood that the effects of auxin on endocytosis and on auxin transport are linked.

Our findings indicate a previously undescribed mode of action of plant hormones, namely the modulation of protein activity by regulating their intracellular trafficking. We have shown that biologically active auxins, but not their biologically inactive analogues nor other plant hormones, negatively regulate endocytosis and the internalization of constitutively cycling proteins from the PM. These auxin-induced changes in the relative rates of endocytosis and exocytosis lead to increased concentrations of PIN auxin efflux regulators at the cell surface. Concomitantly, auxins promote their own efflux by a vesicle recycling-dependent mechanism. These two previously unidentified auxin effects share similar kinetic characteristics, substrate requirements and inhibitor sensitivities. Furthermore, when assessed in parallel *in planta* during the gravitropic response, the same root cells that exhibited increased auxin translocation also displayed a decreased rate of endocytosis. Interestingly, the *sur1* mutant, which has elevated internal auxin concentrations and downregulated endocytosis, also shows increased auxin transport²⁵. All these results strongly indicate that the auxin effects on endocytosis and on its own transport might be functionally linked.

Thus our results show, and provide a mechanistic explanation for, a positive auxin effect on auxin transport rate.

It remains unclear which molecular pathway auxin uses to exhibit its effect on endocytosis. Mutations in the Calossin/Pushover protein BIG render endocytosis partly auxin-resistant. BIG is a single-copy gene present in *Arabidopsis* and other plant and animal genomes. In *Drosophila* the mutations at the corresponding locus lead to multiple defects including altered synaptic transmission and male sterility²⁶; in *Arabidopsis*, big mutations lead to a weak physiological resistance to auxin but also affect other signalling pathways including those for ethylene, cytokinin, gibberellin or light²⁰. These data do not support a direct involvement of the BIG protein in the auxin signalling pathway but rather in more general cellular processes, possibly in some aspect of endocytosis. There is no experimental support for a connection between the BIG-dependent auxin signalling pathway for inhibiting endocytosis and previously characterized SCF^{TIR1}-related auxin signalling for the regulation of gene expression. Thus it is possible that the auxin effect on endocytosis uses a previously unknown and so far molecularly uncharacterized signalling pathway that does not involve the regulation of gene expression.

METHODS

Materials and growth conditions. The following mutants and transgenic plants of *Arabidopsis thaliana* have been described previously: AUX1:HA¹⁴, GNOM-myc², PIN1:GFP⁶, PIN2:GFP¹⁵, DR5rev::GFP⁶, *doc1* (ref. 19), *sur1* (ref. 12), *tir3* (ref. 18), *yucca*¹³, *yucca doc1* double mutant¹⁹, 35S::TMS2 (ref. 11) and EGFP-Q8 (PIP2)²⁷. Experiments were performed on 4-day-old seedlings grown on vertically oriented plates containing *Arabidopsis* medium (AM; half-strength MS agar, 1% sucrose, pH 5.8). Cells of tobacco (*Nicotiana tabacum* L., lines BY-2 and VBI-0) were grown in suspension culture as described elsewhere^{23,24}. Incubation of seedlings with various chemicals was performed in 24-well cell-culture plates in liquid AM medium.

Unless otherwise indicated, the following conditions were used. Pretreatments for 30 min with 5 μ M NAA, 5 μ M IAA plus 400 μ g ml⁻¹ BHT, 5 μ M 2,4-D, 50 μ M 2-NAA, 200 μ M aminocyclopropane carboxylic acid or 10 μ M NAA amide were followed by 90 min of concomitant treatment with one of the above plus 50 μ M BFA. Control treatments contained an equal amount of solvent (dimethylsulphoxide or ethanol). For gravitropism experiments, plants were grown on vertically aligned plates containing a 1-mm layer of AM medium. A gravity stimulus was applied by horizontal positioning of plates; after 60 min, BFA solution was carefully added followed by incubation for 60 min. In controls, gravitational stimulation and BFA treatments were performed in the presence of 10 μ M NPA. All treatments and gravity experiments were performed at least in triplicate, with a minimum of 60 roots evaluated in total in each treatment.

Genetic screen. A mutant screen to isolate plants for resistance to auxin's inhibitory effect on PM protein internalization was performed on the ethyl-methane sulphonate-mutagenized PIN1:GFP⁶ population. Seedlings 5 days old were treated with 30 μ M NAA for 30 min, followed by 30 μ M NAA and 50 μ M BFA for 90 min, and the BFA-induced internalization of PIN1:GFP was analysed with an epifluorescence microscope. From about 3,500 M1 families, we identified eight lines that under these conditions showed resistance to NAA (that is, they displayed normal internalization of PIN1:GFP into BFA compartments, as could be observed in the wild type after treatment with BFA without auxin).

Immunolocalizations. Whole-mount immunofluorescence preparations²⁸ and antibody staining of maize tissue sections⁹ were performed as described. The rabbit anti-PIN1 polyclonal antiserum was raised against amino-acid residues 288–452 of PIN1 protein and was used previously for PIN1 localization in tissue sections⁶. For whole-mount immunolocalization in roots, immunoglobulins from the crude serum were precipitated by saturated (NH₄)₂SO₄ solution (2:1) and dialysed against PBS. The purified fraction was diluted 1:400. The anti-PIN2 antibody was provided by C. Luschnig and was used at a dilution of 1:400. Other antibodies were diluted as follows: anti-PIN4 (1:400)²⁹, anti-GFP (1:300; Molecular Probes), 9E10 anti-Myc (1:600; Santa Cruz), anti-TLG2a (1:200; Rosebiotech), anti-AtSec12 (1:50; Rosebiotech), anti-ARF1 (1:1,000)³⁰, anti-At γ -COP (1:1,000)³ and anti-PM-ATPase (1:1,000)². Fluorescein isothiocyanate-conjugated and CY3-conjugated anti-rabbit secondary antibodies (Dianova) were diluted 1:200 and 1:600, respectively.

Uptake and accumulation experiments. Auxin accumulation experiments in suspension-cultured VBI-0 and BY-2 cells were performed as described previously²⁴. The FM1-43 uptake experiments¹⁶ were performed with BY-2 cells equilibrated in 2,4-D-free medium for 24 h. The measured signal was

normalized to the value of control cells at 4 °C and data were expressed relative to the 26 °C control sample. Uptake experiments with FM4-64 (Molecular Probes) in *Arabidopsis* were performed in 5-day-old seedlings, using a 5-min incubation with 1:500 dilution in AM medium.

Quantitative confocal microscopy. Quantitative confocal microscopy evaluation of the PIN2 signal was performed with Leica LCS quantification software. The scans were performed with identical microscope and laser settings for all experiments. Analysed cells on scans were selected interactively from the cortex of the same root region. Cell-border-associated and internal fluorescence were quantified separately from at least 80 cells for each treatment (10 µM NAA or 50 µM BFA for 90 min). In each experiment, inhibitors of transcription (20 µM cordycepin) and/or protein synthesis (50 µM cycloheximide) were applied to exclude any effects on PIN2 expression. Statistical significance was evaluated with Student's *t*-test.

Electron microscopy. Treatments (10 µM NAA and/or 50 µM BFA for 90 min) and ultrastructure analysis of chemically fixed root sections were performed exactly as described².

Received 1 February; accepted 13 April 2005.

- Royle, S. & Murrell-Lagnado, R. Constitutive cycling: a general mechanism to regulate cell surface proteins. *BioEssays* **25**, 39–46 (2003).
- Geldner, N., Friml, J., Stierhof, Y.-D., Jürgens, G. & Palme, K. Auxin transport inhibitors block PIN1 cycling and vesicle trafficking. *Nature* **413**, 425–428 (2001).
- Geldner, N. *et al.* The *Arabidopsis* GNOM ARF-GEF mediates endosomal recycling, auxin transport, and auxin-dependent plant growth. *Cell* **112**, 219–230 (2003).
- Friml, J., Wisniewska, J., Benková, E., Mendgen, K. & Palme, K. Lateral relocation of auxin efflux regulator AtPIN3 mediates tropism in *Arabidopsis*. *Nature* **415**, 806–809 (2002).
- Friml, J. *et al.* Efflux-dependent auxin gradients establish the apical-basal axis of *Arabidopsis*. *Nature* **426**, 147–153 (2003).
- Benková, E. *et al.* Local, efflux-dependent auxin gradients as a common module for plant organ formation. *Cell* **115**, 591–602 (2003).
- Gray, W. R., Kepinski, S., Rouse, D., Leyser, O. & Estelle, M. Auxin regulates SCF^{TIR1}-dependent degradation of AUX/IAA proteins. *Nature* **414**, 271–276 (2001).
- Friml, J. Auxin transport—shaping the plant. *Curr. Opin. Plant Biol.* **6**, 7–12 (2003).
- Baluška, F. *et al.* F-actin-dependent endocytosis of cell wall pectins in meristematic root cells. Insights from brefeldin A-induced compartments. *Plant Physiol.* **130**, 422–431 (2002).
- Von Gadow, A., Joubert, E. & Hansmann, C. F. Comparison of the antioxidant activity of aspalathin with that of other plant phenols of rooibos tea (*Aspalathus linearis*), alpha-tocopherol, BHT, and BHA. *J. Agric. Food Chem.* **45**, 632–638 (1997).
- Karlin-Neumann, G., Brusslan, J. & Tobin, E. Phytochrome control of the *tms2* gene in transgenic *Arabidopsis*: a strategy for selecting mutants in the signal transduction pathway. *Plant Cell* **3**, 573–582 (1991).
- Boerjan, W. *et al.* Superroot, a recessive mutation in *Arabidopsis*, confers auxin overproduction. *Plant Cell* **7**, 1405–1419 (1995).
- Zhao, Y. *et al.* A role for flavin monooxygenase-like enzymes in auxin biosynthesis. *Science* **291**, 306–309 (2001).
- Swarup, R. *et al.* Localisation of the auxin permease AUX1 suggests two functionally distinct hormone transport pathways operate in the *Arabidopsis* root apex. *Genes Dev.* **15**, 2648–2653 (2001).
- Xu, J. & Scheres, B. Dissection of *Arabidopsis* ADP-RIBOSYLATION FACTOR 1 function in epidermal cell polarity. *Plant Cell* **17**, 525–536 (2005).
- Emans, N., Zimmermann, S. & Fischer, R. Uptake of a fluorescent marker in plant cells is sensitive to brefeldin A and wortmannin. *Plant Cell* **14**, 71–86 (2002).
- Delbarre, A., Muller, P., Imhoff, V. & Guern, J. Comparison of mechanisms controlling uptake and accumulation of 2,4-dichlorophenoxy acetic acid, naphthalene-1-acetic acid, and indole-3-acetic acid in suspension-cultured tobacco cells. *Planta* **198**, 532–541 (1996).
- Ruegger, M. *et al.* Reduced naphthylphthalamic acid binding in the *tir3* mutant of *Arabidopsis* is associated with a reduction in polar auxin transport and diverse morphological defects. *Plant Cell* **9**, 745–757 (1997).
- Gil, P. *et al.* BIG: a calossin-like protein required for polar auxin transport in *Arabidopsis*. *Genes Dev.* **15**, 1985–1997 (2001).
- Kanyuka, K. *et al.* Mutations in the huge *Arabidopsis* gene BIG affect a range of hormone and light responses. *Plant J.* **35**, 57–70 (2003).
- Berleth, T. & Sachs, T. Plant morphogenesis: long-distance coordination and local patterning. *Curr. Opin. Plant Biol.* **4**, 57–62 (2001).
- Morris, D. A. Transmembrane auxin carrier systems—dynamic regulators of polar auxin transport. *Plant Growth Regul.* **32**, 161–172 (2000).
- Petrášek, J., Elčknér, M., Morris, D. A. & Zažímalová, E. Auxin efflux carrier activity and auxin accumulation regulate cell division and polarity in tobacco cells. *Planta* **216**, 302–308 (2002).
- Petrášek, J. *et al.* Do phytochromes inhibit auxin efflux by impairing vesicle traffic? *Plant Physiol.* **131**, 254–263 (2003).
- Delarue, M., Muller, P., Bellini, C. & Delbarre, A. Increased auxin efflux in the IAA-overproducing *sur1* mutant of *Arabidopsis thaliana*: a mechanism of reducing auxin levels? *Physiol. Plant.* **107**, 120–127 (1999).
- Richards, S., Hillman, T. & Stern, M. Mutations in the *Drosophila* pushover gene confer increased neuronal excitability and spontaneous synaptic vesicle fusion. *Genetics* **142**, 1215–1223 (1996).
- Cutler, S., Ehrhardt, D., Griffiths, J. & Somerville, C. Random GFP::cDNA fusions enable visualisation of subcellular structures in cells of *Arabidopsis* at a high frequency. *Proc. Natl Acad. Sci. USA* **97**, 3718–3723 (2000).
- Friml, J., Benková, E., Mayer, U., Palme, K. & Muster, G. Automated whole-mount localization techniques for plant seedlings. *Plant J.* **34**, 115–124 (2003).
- Friml, J. *et al.* AtPIN4 mediates sink-driven auxin gradients and root patterning in *Arabidopsis*. *Cell* **108**, 661–673 (2002).
- Pimpl, P. *et al.* In situ localization and *in vitro* induction of plant COPII-coated vesicles. *Plant Cell* **12**, 2219–2236 (2000).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank C. Bellini, M. Bennett, M. Estelle, M. Grebe, W. Michalke, D. Robinson and Y. Zhao for sharing material, and E. Benková, P. Brewer, J. Eder, J. Malbeck, C. Oecking, M. Sauer, H. Stransky and D. Weijers for technical assistance and discussions. This work was supported by the Volkswagenstiftung (J.F. and T.P.), the F. Ebert Stiftung (J.K.-V.), the Deutsche Forschungsgemeinschaft (N.G., G.J. and Y.-D. S.), the Grant Agency of the Academy of Sciences of the Czech Republic (J.P. and E.Z.) and the Royal Society of London (D.A.M.).

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to J.F. (jiri.friml@zmbp.uni-tuebingen.de).

The conserved protein DCN-1/Dcn1p is required for cullin neddylation in *C. elegans* and *S. cerevisiae*

Thimo Kurz^{1,2}, Nurhan Özlü³, Fabian Rudolf¹, Sean M. O'Rourke², Brian Luke¹, Kay Hofmann⁴, Anthony A. Hyman³, Bruce Bowerman² & Matthias Peter¹

SCF-type E3 ubiquitin ligases are multi-protein complexes required for polyubiquitination and subsequent degradation of target proteins by the 26S proteasome¹. Cullins, together with the RING-finger protein Rbx1, form the catalytic core of the ligase, and recruit the substrate-recognition module^{1–4}. Cycles of covalent modification of cullins by the ubiquitin-like molecule Nedd8 (neddylation)⁵ and removal of Nedd8 by the COP9 signalosome (deneddylation) positively regulate E3 ligase activity^{6,7}. Here we report the identification and analysis of a widely conserved protein that is required for cullin neddylation in the nematode *Caenorhabditis elegans* and the yeast *Saccharomyces cerevisiae*. *C. elegans* DCN-1 and *S. cerevisiae* Dcn1p (defective in cullin neddylation) are characterized by a novel UBA-like ubiquitin-binding domain and a DUF298 domain of unknown function. Consistent with their requirements for neddylation, DCN-1 and Dcn1p directly bind Nedd8 and physically associate with cullins in both species. Moreover, overexpression of Dcn1p in yeast results in the accumulation of Nedd8-modified cullin Cdc53p. Both *in vivo* and *in vitro* experiments indicate that Dcn1p does not inhibit deneddylation of Cdc53p by the COP9 signalosome, but greatly increases the kinetics of the neddylation reaction.

In *C. elegans* the cullin CUL-3 forms a ligase that is required to degrade MEI-1 (ref. 8), a subunit of the microtubule-severing complex katanin⁹. *C. elegans* katanin is active during meiosis¹⁰ but is inactivated by the CUL-3 ligase before mitosis, to allow for the stable assembly and positioning of the first mitotic spindle in the one-cell zygote^{7,11,12}. Loss of the CUL-3 ligase leads to abnormally high levels of katanin during mitosis and associated defects in microtubule-dependent processes.

To identify new proteins involved in CUL-3-dependent degradation of MEI-1/katanin, we screened time-lapse movies of embryonic cell divisions deposited in databases of large-scale *C. elegans* RNA interference (RNAi) screens^{12,13} for defects resembling those caused by loss of CUL-3 function. In the publicly accessible database of the chromosome III RNAi screen¹², we identified a new gene that we named *dcn-1* (open reading frame H38K22.2), which is required for the post-meiotic degradation of MEI-1. In embryos depleted of DCN-1 by RNA interference (hereafter referred to as *dcn-1(RNAi)* embryos), MEI-1 tagged with green fluorescent protein (GFP) accumulated on the mitotic spindle, and endogenous MEI-1 was present at higher levels (Fig. 1a, d; see also Supplementary Videos 1 and 2). Consequently, astral microtubules appeared shorter, and mitotic spindles were mis-positioned (Fig. 1b, c). These mitotic spindle defects in *dcn-1(RNAi)* embryos were rescued by simultaneous reduction of MEI-1 function (Fig. 1c). We conclude that DCN-1 acts with, or in parallel, to the

CUL-3/MEL-26 (where MEL-26 is the substrate-specific adapter of the CUL-3 ligase) E3 ligase to downregulate MEI-1/katanin.

C. elegans DCN-1 is a novel 295-amino-acid protein with a carboxy-terminal DUF298 domain of unknown function (PF03556) (Fig. 2a). This domain is evolutionarily conserved, with a single gene encoding family members in *C. elegans*, *S. cerevisiae* and *Schizosaccharomyces pombe*, and a limited number of homologues in other eukaryotes (Fig. 2b; see also Supplementary Fig. S1). Database searches with generalized profiles constructed from the amino termini of representative DUF298 proteins also revealed a significant similarity ($P < 0.001$) to a number of non-DUF298 proteins. These matches include the human Fas-associated factor FAF1 and the TNF-receptor-associated TTRAP protein, as well as the p97/Cdc48 cofactor p47/Shp1 from several organisms. Inclusion of these proteins in a subsequent iteration of profile searches¹⁴ yielded highly significant matches ($P < 0.001$) to a number of established UBA domains (Supplementary Fig. S2). UBA domains have been found in a variety of proteins linked to the ubiquitin pathway¹⁴, where they have been shown to interact directly with ubiquitin¹⁵. Indeed, purified full-length DCN-1, as well as the UBA-like domain by itself, strongly interacted with immobilized ubiquitin (Fig. 2c), with an affinity significantly stronger than that of the human p47 UBA domain. Interestingly, full-length DCN-1 also directly interacted with the ubiquitin-like protein Nedd8 (Fig. 2c). In contrast to its interaction with ubiquitin, the DCN-1 UBA-like domain by itself only weakly interacted with Nedd8, suggesting that the UBA-like domain is specific for ubiquitin, and that Nedd8 binding is probably mediated by a different part of DCN-1. We obtained similar results with the *S. cerevisiae* DCN-1 homologue Dcn1p (YLR128W), suggesting that both ubiquitin- and Nedd8-binding of these proteins is evolutionarily conserved. Although many of the DUF298-domain-containing proteins contain an N-terminal UBA-like domain, it is absent in some of the DCN-1 homologues in other organisms (Fig. 2b). Notably, *C. elegans* DCN-1 is expressed in two different splice forms, with the UBA-like domain absent in the shorter one (Supplementary Fig. S3), suggesting that DCN-1 and DCN-1-like proteins may perform functions independent of their ubiquitin-binding properties and more directly related to Nedd8 binding.

Because DCN-1 and Dcn1p both bind Nedd8, and *dcn-1(RNAi)* results in defects similar to those observed in mutant *C. elegans* embryos lacking the neddylation pathway, we investigated whether DCN-1 and Dcn1p are involved in cullin neddylation. In *C. elegans* embryos expressing a GFP-tagged version of DCN-1, the GFP signal was highest in the nuclei of early embryonic cells but was also present at lower levels in the cytoplasm (Supplementary Fig. S4). We observed an identical expression pattern using anti-DCN-1 antibodies and indirect immunofluorescence to stain fixed embryos

¹Institute of Biochemistry, ETH Hönggerberg, 8093 Zürich, Switzerland. ²Institute of Molecular Biology, University of Oregon, Eugene, Oregon 97403, USA. ³Max Planck Institute of Molecular Cell Biology and Genetics, 01307 Dresden, Germany. ⁴Memorec Biotec GmbH, 50829 Köln, Germany.

(Supplementary Fig. S4). This distribution of DCN-1 in *C. elegans* embryonic cells is identical to that previously reported for *C. elegans* Nedd8 and for the COP9 signalosome subunit CSN-5 (refs 7, 8), consistent with a role for DCN-1 in cullin neddylation. We next used western blots to compare the status of CUL-3 neddylation in extracts prepared from wild-type and *dcn-1(RNAi)* embryos. Interestingly, endogenous CUL-3 was mainly present in its un-neddyated form in *dcn-1(RNAi)* extracts, comparable to embryos lacking the Nedd8 E1

activating enzyme RFL-1 (Fig. 3a). In budding yeast, deletion mutations of the genes affecting the neddylation status of Cdc53p (Cul1) are not lethal, but they lower the restrictive temperature for conditional alleles of Cdc53p and other components of this E3 ligase^{16,17}. Accordingly, deletion of *S. cerevisiae* DCN1 in yeast carrying the temperature-sensitive *cdc53-1* allele resulted in lethality at lower temperatures compared with *cdc53-1* cells with wild-type DCN1 (Fig. 3c); neddyated Cdc53p was barely detectable in *dcn1Δ*

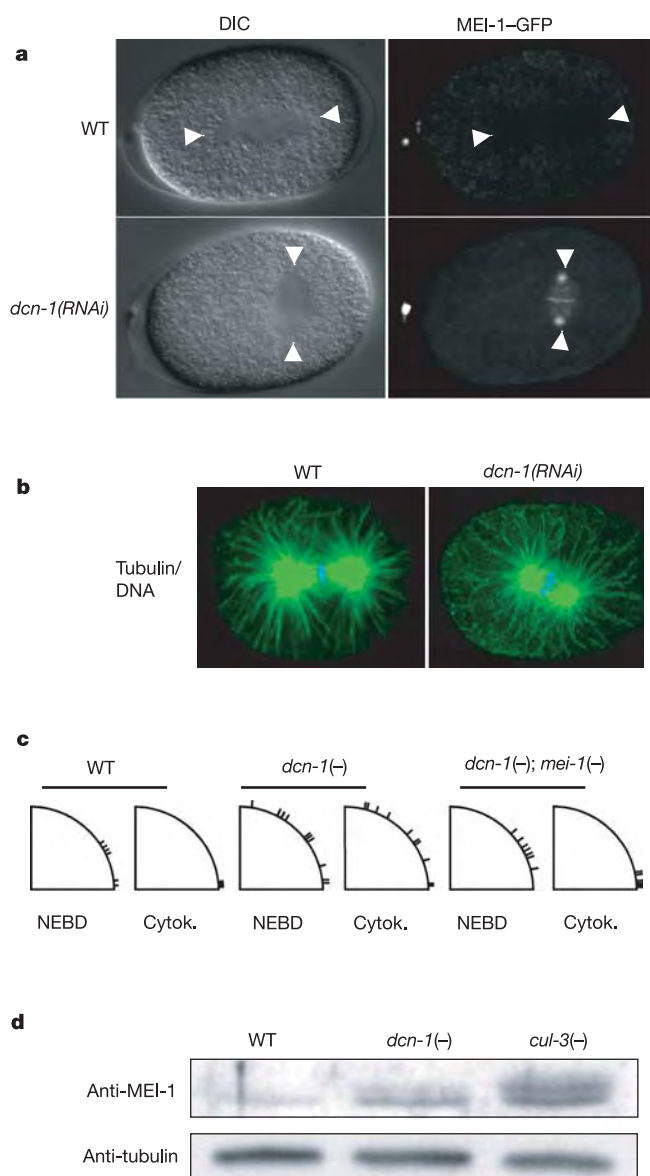


Figure 1 | *dcn-1(RNAi)* embryos stabilize MEI-1 protein. **a**, Nomarski and MEI-1-GFP micrographs of wild-type (top row) and *dcn-1(RNAi)* (bottom row) embryos during the first cell division. Loss of DCN-1 results in spindle orientation defects (arrowheads) and ectopic MEI-1 localization to the centrosomes and the mitotic apparatus (bottom-right panel). **b**, Confocal micrographs of α -tubulin and DNA in one-cell embryos fixed for indirect immunofluorescence. Compared with wild-type spindles, *dcn-1(RNAi)* spindles are shorter, mis-oriented and many astral microtubules terminate before reaching the cortex. **c**, Spindle orientation in multiple wild-type, *dcn-1(RNAi)* and *dcn-1(RNAi); mei-1(RNAi)* embryos, at nuclear envelope breakdown (NEBD) and during cytokinesis (Cytok.). Alignment along the anterior–posterior axis corresponds to 0°. Each black bar represents one embryo. Simultaneous reduction of MEI-1 function restores a wild-type spindle orientation in *dcn-1(RNAi)* embryos. **d**, MEI-1 levels in extracts prepared from wild-type, *dcn-1(RNAi)* and *cul-3(RNAi)* embryos.

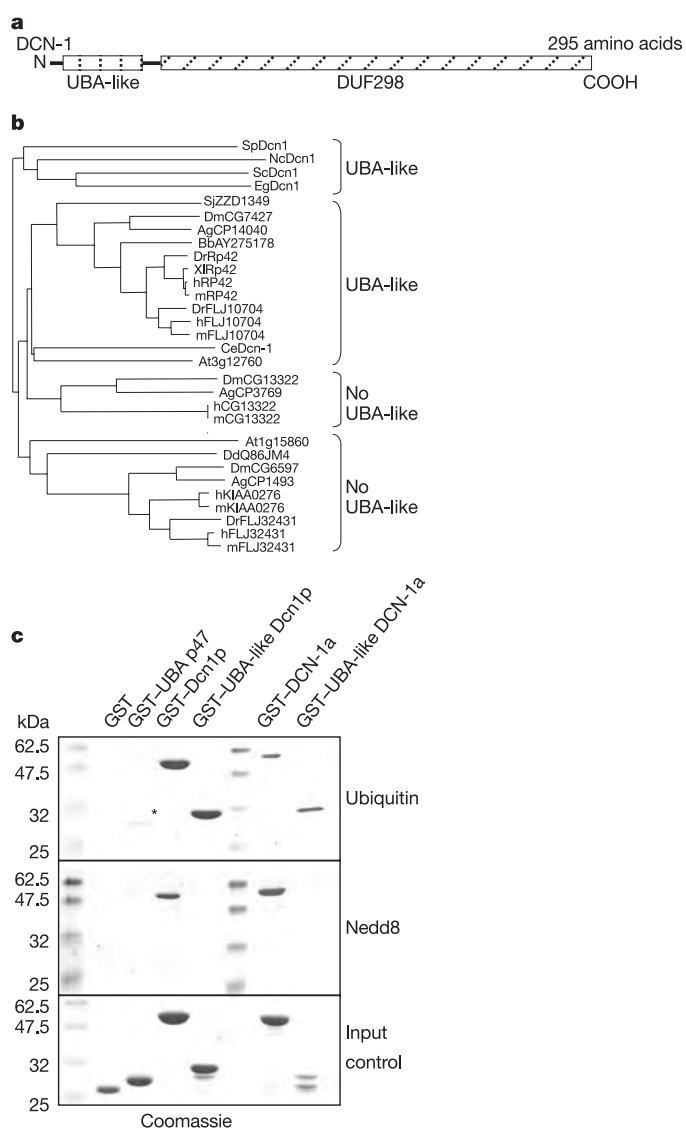


Figure 2 | DCN-1 and Dcn1p are evolutionarily and functionally conserved and bind ubiquitin and Nedd8. **a**, *C. elegans* DCN-1 is characterized by a C-terminal DUF298 domain of unknown function (PF03556) and a novel N-terminal UBA-like domain. **b**, Neighbour-joining tree of representative DUF298 domains. Only two of the major subgroups are found associated with UBA-like domains. Ag, *Anopheles gambiae*; At, *Arabidopsis thaliana*; Bb, *Branchiostoma belcheri*; Dd, *Dictyostelium discoideum*; Dm, *Drosophila melanogaster*; Dr, *Danio rerio*; Eg, *Eremothecium gossypii*; h, *Homo sapiens*; m, *Mus musculus*; Nc, *Neurospora crassa*; Sc, *S. cerevisiae*; Sj, *Schistosoma japonicum*; Sp, *S. pombe*; Xl, *Xenopus laevis*. **c**, DCN-1a and Dcn1p directly bind to ubiquitin and Nedd8. A tenfold molar excess of full-length and the indicated domains of DCN-1 and Dcn1p were expressed as GST fusion proteins, and assayed for direct binding to immobilized ubiquitin or Nedd8. The known ubiquitin-binding UBA domain of human p47 (asterisk) served as a positive control. The bottom panel shows the protein input (10% for ubiquitin-binding assay; 2.5% for Nedd8-binding assay).

cells (Fig. 3b). Thus, DCN-1 and Dcn1p function is required for cullin neddylation in two highly diverged species.

Because Dcn1p was found to physically interact with Cdc53p in a genome-wide two-hybrid screen¹⁸, we reasoned that Dcn1p might promote Nedd8 conjugation in part by binding to the cullin. We first verified the interaction of Dcn1p with the CUL1 homologue Cdc53p using a LexA-based yeast two-hybrid system (Fig. 3d). Notably, this interaction is conserved, because endogenous *C. elegans* CUL-3 in nematode extracts associated with glutathione *S*-transferase (GST)–DCN-1, but not with control GST, in pull-down experiments (Fig. 3e). Moreover, the interaction of DCN-1 and Dcn1p with two different types of cullins indicates that they may act on other cullin subclasses. Indeed, the neddylation state of the yeast cullin Rtt101p is also affected by the deletion of *DNC1* (Supplementary Fig. S4), and a large-scale yeast two-hybrid analysis in *Drosophila melanogaster* revealed that the *Drosophila* DUF298 domain proteins CG7427 and CG13322 interact with *Drosophila* Cul5 and Cul4, respectively¹⁹.

Loss of DCN-1/Dcn1p function could result in the accumulation of un-neddyated cullin in either of two ways: a defect in neddylation, or a failure to inhibit deneddylation by the COP9 signalosome. To distinguish between these two possibilities, we deleted both *DCN1* and the core subunit of the signalosome *RR11* in budding yeast. If loss of Dcn1p promotes deneddylation, the double mutant should resemble an *rr1Δ* single mutant, and exclusively accumulate neddylated Cdc53p. In contrast, if Dcn1p is required for neddylation, the double mutant should resemble a *dcn1Δ* single mutant, with reduced levels of neddylated Cdc53p. Notably, we detected both neddylated and un-neddylated Cdc53p in the *rr1Δ dcn1Δ* double mutant (Fig. 4a), implying that Dcn1p is not essential for the

neddylation reaction in the absence of deneddylation activity, but increases its efficiency *in vivo*. Importantly, overexpression of Dcn1p promoted a shift to the neddylated form of Cdc53p (Fig. 4a) and also lowered the permissive temperature of *cdc53-1* temperature-sensitive mutants (Fig. 4b), comparable to the deletion of the deneddylase *RR11* (ref. 20). Together, these observations suggest that Dcn1p may be a limiting factor that promotes cullin neddylation *in vivo*.

To confirm that the Nedd8 modification machinery indeed works less efficiently in cells for which *DCN1* has been deleted, we expressed haemagglutinin (HA)-tagged Cdc53p from the inducible *GAL* promoter in both *rr1Δ* and *rr1Δ dcn1Δ* mutants. After induction of HA-Cdc53p in galactose-containing medium for 1 h, its expression was shut off transcriptionally by the addition of glucose. We then examined the neddylation state of HA-Cdc53p over a period of 3 h after shut off. Unlike endogenous Cdc53p, overexpressed HA-tagged Cdc53p is not completely neddylated in *RR11*-deleted cells (Fig. 4c). However, most of the protein is modified and the levels of neddylated Cdc53p did not change over time (Fig. 4c). In contrast, only a small amount of HA-Cdc53p was modified in *rr1Δ dcn1Δ* double mutants 1 h after induction of the protein. The amount of modified HA-Cdc53p slowly increased over time, but even at 3 h after the shut off, neddylated HA-Cdc53p was less abundant than in *rr1Δ* single mutants (Fig. 4c). Thus, we conclude that Dcn1p greatly promotes cullin neddylation *in vivo*.

We next tested whether Dcn1p can also promote cullin neddylation *in vitro*. Previous work had shown that neddylation can be achieved *in vitro* when the cullin is in a complex with Rbx1 (ref. 21), suggesting that Rbx1 might function as the ligase for Nedd8. As expected, a small fraction of Cdc53p was neddylated *in vitro* in the purified Cdc53p–Rbx1p complex in the presence of Nedd8, the E1

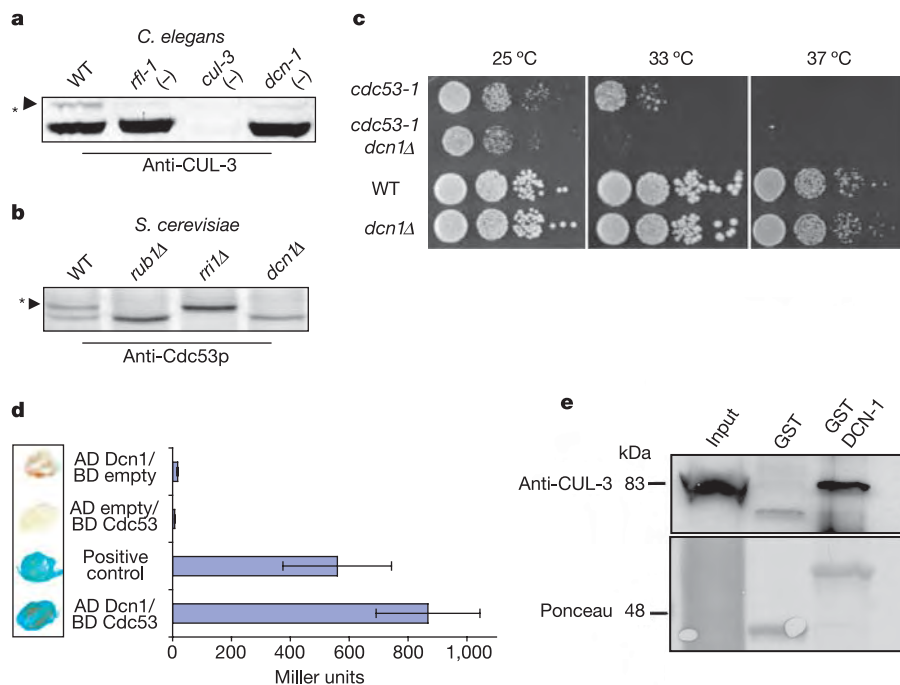


Figure 3 | DCN-1 and Dcn1p proteins are required for Nedd8 modification of cullins in yeast and nematodes, and physically associate with cullins.

a, *C. elegans* extracts prepared from wild-type, Nedd8 E1 activating enzyme mutants (*rrf-1*), *cul-3*(RNAi) and *dcn-1*(RNAi) mutant embryos were subjected to immunoblotting with anti-CUL-3 antibody. Nedd8-modified Cul-3 is marked with an asterisk. **b**, Extracts from wild-type, *rub1Δ*, *rri1Δ* and *dcn1Δ* yeast cells were subjected to immunoblotting with anti-Cdc53p antibody. Deletion of the deneddylase *RR11* resulted in accumulation of Nedd8/Rub1p-modified Cdc53p (asterisk), whereas in *rub1Δ* and *dcn1Δ*

cells only unmodified Cdc53p was detected. **c**, Deletion of *S. cerevisiae* *DCN1* lowers the restrictive temperature of *cdc53-1* conditional mutants. **d**, The interaction of Dcn1p (AD; activation domain) and Cdc53p (BD; binding domain) was determined by a two-hybrid assay by LacZ-reporter activity on filters (left) and quantitatively in liquid assays (right). Error bars indicate s.d. of at least three independent experiments. **e**, GST–DCN-1 and GST were expressed in *E. coli* and incubated with nematode extracts. Immunoblotting was performed using anti-CUL-3 antibody (top panel). The bottom panel shows Ponceau S staining of the same blot.

activating complex APP-BP1-UBA3 and the E2 UbcH12 (Fig. 4d). Importantly, the fraction of neddylation of Cdc53p was significantly increased by the addition of yeast extract expressing GFP-Dcn1p, but to a much lower extent by addition of an extract prepared from *dcn1Δ* cells. Quantification of time course experiments revealed that the presence of GFP-Dcn1p reduced the $t_{1/2}$ of the reaction from 42 min to 2.4 min (Fig. 4e, f), showing that Dcn1p increases the kinetics of the neddylation reaction approximately 20-fold. Interestingly, purified Dcn1p expressed in *Escherichia coli* or SF9 cells was not sufficient to promote neddylation of Cdc53p (data not shown), implying that either a post-translational modification of Dcn1p or additional components in the extract may be involved. It is interesting to note in this regard that even addition of *dcn1Δ* extracts slightly increased the basal neddylation activity of the Cdc53p-Rbx1p complex (Fig. 4d).

Cullin neddylation and deneddylation are highly conserved processes that influence the activity of most SCF-type E3 ligases. The function of cullin neddylation remains elusive, but it has been proposed that the Nedd8 moiety initiates E3 ligase assembly^{22,23} and increases the affinity of the E2 to the ligase²⁴. As with ubiquitination, neddylation of cullins requires an E1 activating and E2 conjugating enzyme; an E3 ligase for Nedd8 conjugation has not been identified, although *in vitro* reconstitution assays suggested that Rbx1 may perform this function²¹. Our results suggest that DCN-1 and Dcn1p may function either as a regulator or part of a Nedd8 E3

ligase. This conclusion is consistent with our finding that the *in vivo* and *in vitro* kinetics of the neddylation reaction are significantly increased by the presence of Dcn1p. Furthermore, overexpression of GFP-Dcn1p in *S. cerevisiae* converts the cullin Cdc53p into its Nedd8-modified form. Finally, DCN-1 and Dcn1p physically associate with cullins in both *C. elegans* and *S. cerevisiae*, and directly bind to Nedd8. However, we do not believe that Dcn1p acts as a Nedd8 E3 ligase by itself, because bacterially or baculovirus-expressed purified Dcn1p was not able to enhance the kinetics of the neddylation reaction *in vitro*. Moreover, *dcn1Δ* extracts mildly stimulated Cdc53p neddylation *in vitro*, further suggesting the presence of additional positive regulators of neddylation.

Although the exact molecular mechanism of DCN-1/Dcn1p function remains unclear, it is likely that direct Nedd8 binding to DCN-1 and Dcn1p has an important role. The Nedd8 E2 Ubc12 has been shown to associate with Rbx1 (ref. 25), and the addition of Rbx1 is necessary to reconstitute neddylation activity *in vitro*²¹. It is possible that DCN-1/Dcn1p and Rbx1 function together to recruit the Nedd8-charged E2 Ubc12 to the cullin and thus facilitate neddylation. Additionally, it is important to emphasize that DCN-1 and Dcn1p contain a strong UBA-like ubiquitin-binding domain at their N termini. How ubiquitin binding may contribute to efficient neddylation is not immediately apparent, and it is tempting to speculate that the ubiquitin- and Nedd8-binding properties of DCN-1 and Dcn1p are required to execute two independent functions.

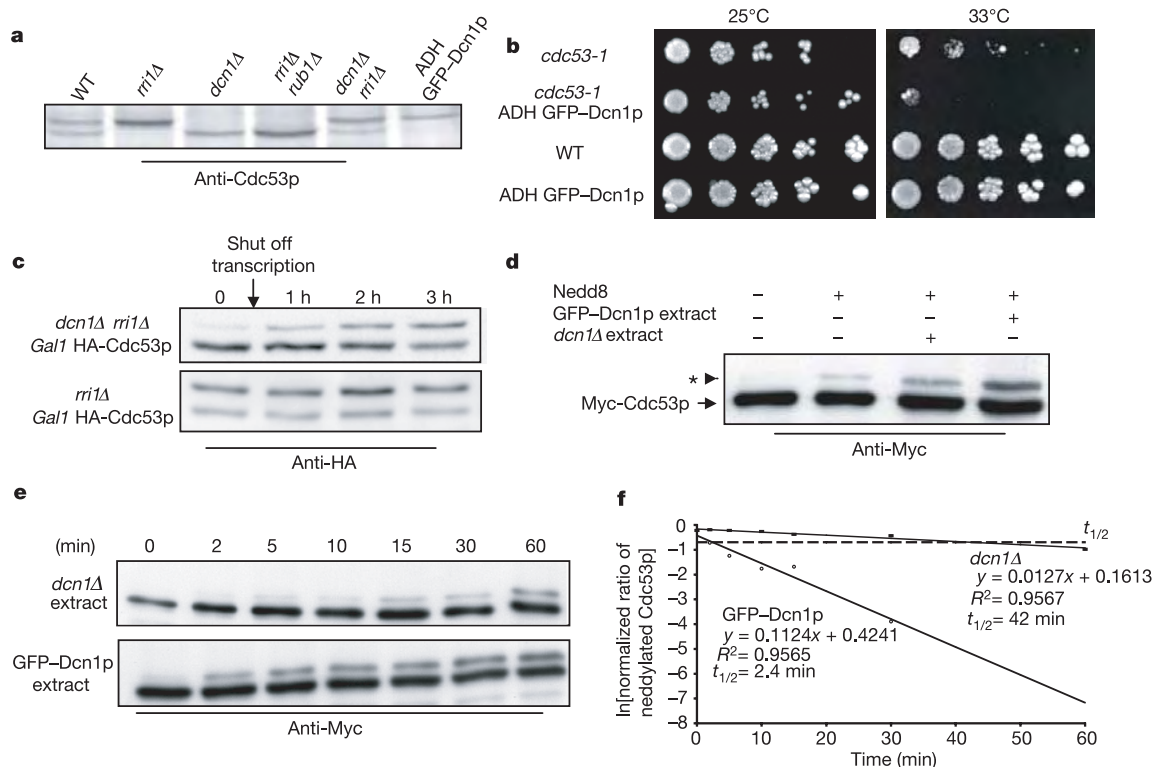


Figure 4 | DCN-1 and Dcn1p catalyse cullin neddylation *in vivo* and *in vitro*.

a, b, The neddylation state of Cdc53p was determined by a western blot of yeast cell extracts prepared from *rri1Δ dcn1Δ* and *rub1Δ rri1Δ* cells. Constitutive ADH overexpression of GFP-Dcn1p results in accumulation of Nedd8-modified Cdc53p (**a**), and lowers the permissive temperature of *cdc53-1* temperature-sensitive mutants (**b**). **c**, The kinetics of Cdc53p neddylation *in vivo* was compared in *dcn1Δ rri1Δ* double (top panel) and *rri1Δ* single mutants (bottom panel) by inducing expression of HA-tagged Cdc53p from the GAL promoter. The transcription of HA-Cdc53p was shut off 1 h after induction, and the neddylation state of Cdc53p was followed for 3 h after shut off. **d**, Dcn1p promotes the neddylation of Cdc53p *in vitro*. Purified Myc-Cdc53p-Rbx1p complex was used as substrate for a 15 min *in*

vitro neddylation reaction containing Nedd8, yeast extract, Nedd8 E1, Nedd8 E2 and ATP. The asterisk marks the neddylated form of Cdc53p.

e, f, *In vitro* neddylation in the presence of *dcn1Δ* or GFP-Dcn1p extract over time. Reactions were started by the simultaneous addition of Nedd8 and yeast extract, and samples were taken at the time points indicated. The fraction of neddylation Myc-Cdc53p at each time point was quantified and plotted as a function of the neddylation state against time. The fraction of neddylation cullin was determined as $-\ln([Nedd-Myc-Cdc53p]_{\infty} - [Nedd-Myc-Cdc53p]_t)$, whereas the 60-min time point for GFP-Dcn1p was chosen as $[Nedd-Myc-Cdc53p]_{\infty}$. The different concentrations for $[Nedd-Myc-Cdc53p]_t$ were normalized to total $[Myc-Cdc53p]$ in each reaction.

METHODS

Strains and manipulations. *C. elegans* Bristol strain N2 was used as wild type. Injection RNAi was performed as described previously⁸ or by feeding nematodes *E. coli* expressing RNA from L4440-derived feeding vectors. To identify new genes involved in MEI-1 degradation, movies deposited in databases from large-scale RNAi screens (<http://www.wormbase.org>) were visually screened for defects resembling *cul-3* loss of function and inactivated by RNAi in GFP–MEI-1-expressing nematodes.

Yeast strains are listed in Supplementary Table 1. Standard yeast growth conditions and genetic manipulations were used²⁶. To test for an enhancement of *cdc53-1* temperature sensitivity, the strains indicated in Figs 2c and 4b were grown to mid-log phase at permissive temperature, and subsequently fivefold serial dilutions of an equal number of cells were spotted onto YPD medium and grown at 25 °C, 33 °C and 37 °C for 2–3 days.

Alignments and sequence analysis. Database and Blast searches were carried out at NCBI (<http://www.ncbi.nlm.nih.gov>), SGD (<http://www.yeastgenome.org>) and Wormbase (<http://www.wormbase.org>). Multiple alignments and general sequence analysis was performed using T-coffee, Boxshade 3.2 and MacVector software. Generalized profile construction and searches were run locally using the *pftools* package, version 2.1. (program available from <ftp://ftp.isrec.isb-sib.ch/sib-isrec/pftools/>). Profiles were constructed using the BLOSUM45 substitution matrix and default penalties of 2.1 for gap opening and 0.2 for gap extension²⁰. A significance threshold of $P < 0.01$, derived from database shuffling analysis, was used as an acceptance criterion for iterative profile refinement²⁷.

Protein extracts, antibodies and immunoblotting. *C. elegans* embryonic extracts and yeast protein extracts were prepared as described previously^{7,28}. The following antibodies were used in this study: anti-MEI-1, anti-CUL-3, anti-DCN-1, anti- α -tubulin (Sigma clone DM1 α), anti-HA (Babco) and anti-Cdc53p (Santa Cruz; yN-18). Secondary antibodies conjugated to horseradish peroxidase were purchased from BioRad and used according to the manufacturers specifications.

Polyclonal antibodies to *C. elegans* DCN-1 were generated by injecting purified GST-fused full-length protein into rabbits. For affinity purifications, GST fusions were cleaved using Precision protease (Amersham Pharmacia Biotech), DCN-1 was coupled to HiTrap-NHS columns (Amersham Pharmacia Biotech) and rabbit sera were affinity purified against coupled columns.

In vivo and in vitro Cdc53p neddylation. Full-length HA-Cdc53p was cloned under the inducible *GAL1,10* promoter and transformed into *rrl1* Δ (YTK46) and *rrl1* Δ *dcn1* Δ (YTK43) yeast. The cells were grown in 2% raffinose to log phase, when HA-Cdc53p expression was induced by the addition of 2% galactose. After 1 h, HA-Cdc53p expression was shut off by addition of 2% glucose, and samples were taken until 3 h after shut off.

For *in vitro* neddylation, the substrate (Myc3-Cdc53–6 $\times$ His-Rbx1 complex) immobilized on Ni-NTA beads was added to the reaction buffer (50 mM Tris-HCl pH 7.5, 10 mM MgCl₂, 2 mM ATP, 50 μ M dithiothreitol (DTT)) containing human Nedd8 E1 (Boston Biochem) and human Nedd8 E2 (Boston Biochem). The reaction was started by the addition of human 6 $\times$ His-Nedd8 (Boston Biochem) or a mixture of 6 $\times$ His-Nedd8 and yeast extract and incubated at 30 °C. The reaction was stopped at the appropriate time by addition of SDS sample buffer. Neddylated Cdc53p was detected by western blotting using anti-Myc antibodies (Gramsch Laboratories). High concentration yeast spheroblast extract was essentially prepared as described previously²⁹.

Ubiquitin- and Nedd8-binding assays. The respective complementary DNAs were cloned into pGEX4T (Amersham) and expressed in BL21 D3 Star cells (Invitrogen). All proteins were purified to homogeneity and concentrated to 1 mM in storage buffer (50 mM KCl, 25 mM Tris pH 7.4 and 5% glycerol). Binding experiments were carried out at 4 °C in 1 ml of binding buffer. Binding buffer contained 150 mM KCl, 50 mM Tris pH 8.0, 2 mM DTT, 5% glycerol and 0.1% Triton X-100 for the *S. cerevisiae* proteins and 150 mM KCl, 50 mM Tris pH 7.4, 2 mM DTT, 5% glycerol and 0.1% Triton X-100 for the *C. elegans* proteins. For one reaction, 5 μ l (50 μ g) of agarose-immobilized ubiquitin or Nedd8 (Boston Biochem) were extensively washed with binding buffer, incubated with 50 μ l (50 nmol) of each protein for 3 h at 4 °C on a wheel, subsequently washed four times for 15 min with 1 ml binding buffer, and eluted with SDS buffer. Bound proteins were detected on a 10% SDS gel by Coomassie staining.

Received 10 February; accepted 15 April 2005.

- Deshaies, R. J. SCF and Cullin/Ring H2-based ubiquitin ligases. *Annu. Rev. Cell Dev. Biol.* **15**, 435–467 (1999).
- Kile, B. T. et al. The SOCS box: a tale of destruction and degradation. *Trends Biochem. Sci.* **27**, 235–241 (2002).

- Pintard, L. et al. The BTB protein MEL-26 is a substrate-specific adaptor of the CUL-3 ubiquitin-ligase. *Nature* **425**, 311–316 (2003).
- Furukawa, M., Ohta, T. & Xiong, Y. Activation of UBC5 ubiquitin-conjugating enzyme by the RING finger of ROC1 and assembly of active ubiquitin ligases by all cullins. *J. Biol. Chem.* **277**, 15758–15765 (2002).
- Pan, Z. Q., Kentsis, A., Dias, D. C., Yamoah, K. & Wu, K. Nedd8 on cullin: building an expressway to protein destruction. *Oncogene* **23**, 1985–1997 (2004).
- Lyapina, S. et al. Promotion of NEDD-CUL1 conjugate cleavage by COP9 signalosome. *Science* **292**, 1382–1385 (2001).
- Pintard, L. et al. Neddylation and deneddylation of CUL-3 is required to target MEI-1/katanin for degradation at the meiosis-to-mitosis transition in *C. elegans*. *Curr. Biol.* **13**, 911–921 (2003).
- Kurz, T. et al. Cytoskeletal regulation by the Nedd8 ubiquitin-like protein modification pathway. *Science* **295**, 1294–1298 (2002).
- McNally, F. J. & Vale, R. D. Identification of katanin, an ATPase that severs and disassembles stable microtubules. *Cell* **75**, 419–429 (1993).
- Srayko, M., Buster, D. W., Bazirgan, O. A., McNally, F. J. & Mains, P. E. MEI-1/MEI-2 katanin-like microtubule severing activity is required for *Caenorhabditis elegans* meiosis. *Genes Dev.* **14**, 1072–1084 (2000).
- Dow, M. R. & Mains, P. E. Genetic and molecular characterization of the *Caenorhabditis elegans* gene, mel-26, a postmeiotic negative regulator of mei-1, a meiotic-specific spindle component. *Genetics* **150**, 119–128 (1998).
- Gonczy, P. et al. Functional genomic analysis of cell division in *C. elegans* using RNAi of genes on chromosome III. *Nature* **408**, 331–336 (2000).
- Fraser, A. G. et al. Functional genomic analysis of *C. elegans* chromosome I by systematic RNA interference. *Nature* **408**, 325–330 (2000).
- Bucher, P., Karplus, K., Moeri, N. & Hofmann, K. A flexible motif search technique based on generalized profiles. *Comput. Chem.* **20**, 3–23 (1996).
- Meyer, H. H., Wang, Y. & Warren, G. Direct binding of ubiquitin conjugates by the mammalian p97 adaptor complexes, p47 and Ufd1-Npl4. *EMBO J.* **21**, 5645–5652 (2002).
- Liakopoulos, D., Doenges, G., Matuschewski, K. & Jentsch, S. A novel protein modification pathway related to the ubiquitin system. *EMBO J.* **17**, 2208–2214 (1998).
- Lammer, D. et al. Modification of yeast Cdc53p by the ubiquitin-related protein rub1p affects function of the SCF/Cdc4 complex. *Genes Dev.* **12**, 914–926 (1998).
- Uetz, P. et al. A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*. *Nature* **403**, 623–627 (2000).
- Giot, L. et al. A protein interaction map of *Drosophila melanogaster*. *Science* **302**, 1727–1736 (2003).
- Cope, G. A. et al. Role of predicted metalloprotease motif of Jab1/Csn5 in cleavage of Nedd8 from Cul1. *Science* **298**, 608–611 (2002).
- Morimoto, M., Nishida, T., Nagayama, Y. & Yasuda, H. Nedd8-modification of Cul1 is promoted by Roc1 as a Nedd8–E3 ligase and regulates its stability. *Biochem. Biophys. Res. Commun.* **301**, 392–398 (2003).
- Zheng, J. et al. CAND1 binds to unneddyated CUL1 and regulates the formation of SCF ubiquitin E3 ligase complex. *Mol. Cell* **10**, 1519–1526 (2002).
- Liu, J., Furukawa, M., Matsumoto, T. & Xiong, Y. NEDD8 modification of CUL1 dissociates p120(CAND1), an inhibitor of CUL1–SKP1 binding and SCF ligases. *Mol. Cell* **10**, 1511–1518 (2002).
- Kawakami, T. et al. NEDD8 recruits E2-ubiquitin to SCF E3 ligase. *EMBO J.* **20**, 4003–4012 (2001).
- Dharmasiri, S., Dharmasiri, N., Hellmann, H. & Estelle, M. The RUB/Nedd8 conjugation pathway is required for early development in *Arabidopsis*. *EMBO J.* **22**, 1762–1770 (2003).
- Guthrie, C. & Fink, G. R. in *Guide to Yeast Genetics and Molecular Biology* (eds Abelson, J. N. & Simon, M. I.) (Academic, San Diego, California, 1991).
- Hofmann, K. Sensitive protein comparisons with profiles and hidden Markov models. *Brief. Bioinform.* **1**, 167–178 (2000).
- Jaquenoud, M., Gulli, M. P., Peter, K. & Peter, M. The Cdc42p effector Gic2p is targeted for ubiquitin-dependent degradation by the SCFGrr1 complex. *EMBO J.* **17**, 5360–5373 (1998).
- Verma, R., Chi, Y. & Deshaies, R. J. Cell-free ubiquitination of cell cycle regulators in budding yeast extracts. *Methods Enzymol.* **283**, 366–376 (1997).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank M. Srayko for providing DCN-1 cDNA; M. Tyers and X. Tang for purified Cdc53p–Rbx1p complex; S. Rybina for help with antibody purification; and J. C. Labbé & L. Pintard for critical reading of the manuscript. T.K. was supported by an American Heart Association Pre-doctoral Fellowship and an EMBO Long-Term Fellowship, S.M.O. by the Damon Runyon Cancer Research Foundation, B.B. by the NIH, and M.P. by the ETHZ and the Swiss National Science Foundation.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to B.B. (bbowerman@molbio.uoregon.edu) or M.P. (matthias.peter@bc.biol.ethz.ch).

LETTERS

Global histone modification patterns predict risk of prostate cancer recurrence

David B. Seligson^{1*}, Steve Horvath^{2,3*}, Tao Shi^{2,3}, Hong Yu¹, Sheila Tze¹, Michael Grunstein⁴
& Siavash K. Kurdistani⁴

Aberrations in post-translational modifications of histones have been shown to occur in cancer cells but only at individual promoters¹; they have not been related to clinical outcome. Other than being targeted to promoters, modifications of histones, such as acetylation and methylation of lysine and arginine residues, also occur over large regions of chromatin including coding regions and non-promoter sequences, which are referred to as global histone modifications². Here we show that changes in global levels of individual histone modifications are also associated with cancer and that these changes are predictive of clinical outcome. Through immunohistochemical staining of primary prostatectomy tissue samples, we determined the percentage of cells that stained for the histone acetylation and dimethylation of five residues in histones H3 and H4. Grouping of samples with similar patterns of modifications identified two disease subtypes with distinct risks of tumour recurrence in patients with low-grade prostate cancer. These histone modification patterns were predictors of outcome independently of tumour stage, preoperative prostate-specific antigen levels, and capsule invasion. Thus, widespread changes in specific histone modifications indicate previously undescribed molecular heterogeneity in prostate cancer and might underlie the broad range of clinical behaviour in cancer patients.

Cancer of the prostate shows a heterogeneous clinical behaviour, from indolent to highly aggressive, and is the second leading cause of cancer deaths in men in the United States³. Present biomarkers including preoperative prostate-specific antigen (PSA) and biopsy Gleason score⁴ (a measure of tumour differentiation, scored from 2 to 10 with increasing degree of dedifferentiation) have not proved to be accurate predictors of clinical outcome⁵. The lack of prognostic markers is even more pressing in younger men with asymptomatic low-grade (Gleason score less than 7) tumours who are being increasingly identified by prevalent screening of PSA levels, but it is unclear how aggressively they should be treated^{6,7}. Improved prognostic markers are therefore needed.

Enzymes that modify histones show altered activity in cancer. For instance, missense mutations of p300 histone acetyltransferases and loss of heterozygosity at the p300 locus are associated with colorectal and breast cancers and with glioblastomas^{8–10}. The consequence of the altered activity of histone-modifying enzymes has so far been linked to inappropriate expression of few genes that might have a function in tumour biology. For instance, p300 is involved in androgen receptor transactivation, with a potentially important function in the progression of prostate cancer¹¹. However, in addition to being targeted to promoters, these enzymes also affect most

nucleosomes throughout the genome independently of apparent sequence-specific DNA-binding proteins^{2,12,13}. Furthermore, the histone-modifying enzymes possess a high degree of substrate specificity that differentiates between both the histone subtypes and the individual side chains within each histone^{14,15}. Thus, individual residues may be modified globally to various extents, reflecting the selective but widespread activity of the histone-modifying enzymes.

To determine the global levels of individual histone modifications in tissues obtained from patients, we combined immunohistochemistry, a method for detecting the presence of specific antigens in cells, with tissue microarrays (TMAs), for high-throughput analysis of many tissue samples¹⁶. We analysed the levels of acetylated (Ac) H3 Lys 9 (K9), K18 and H4 K12, and of dimethylated (diMe) H4 Arg 3 (R3) and H3 K4, using highly specific antibodies¹⁵ (Supplementary Fig. S1), on 183 primary prostate cancer tissues. The level of staining was assessed independently by two pathologists, who were blinded to all clinico-pathological variables. Here the global level of staining refers to the percentage of cells within each tissue sample that stained positively for a given antibody. For instance, Fig. 1a–d shows representative staining of four tissue samples, two each for H3 K18Ac (Fig. 1a, b) and H4 R3diMe (Fig. 1c, d) on tissue arrays. The cells with brown nuclei are considered positively stained. The unstained cells may still contain the modifications at certain genomic loci but their levels are below the detection limits, signifying that bulk histone modifications are considerably decreased in these cells. Immunostaining therefore reveals the presence or absence of global histone modifications in primary tissues.

To assess the differences in staining for the five antibodies, we plotted the frequencies (*y* axis) of tissue samples in which the indicated percentage cell staining (*x* axis) were observed for each modification (Fig. 1e). Acetylations of H3 K9 and K18 had very similar distributions, with more than 65% of samples showing 90–100% staining of the tumour cells. In contrast, only 16% of samples showed 90–100% staining for H4 K12; little or no acetylation was detected in about 24% of samples. Dimethylation of H3 K4 and H4 R3 showed broader distributions, with more than 60% of samples staining between 20% and 80% of cells. These data indicate that the levels of histone modifications differ considerably between individual tissues. We show below that these differences are important for defining groups of patients with distinct clinical outcomes.

We first wished to determine the relationship between levels of the five histone modifications and degree of differentiation (grade) for all samples. Of the five modifications, H3 K18Ac ($r = 0.28$, $P = 9.6 \times 10^{-5}$), H3 K4diMe ($r = 0.22$, $P = 3.4 \times 10^{-3}$),

¹Department of Pathology and Laboratory Medicine, David Geffen School of Medicine, University of California, Los Angeles, California 90095, USA. ²Department of Human Genetics, David Geffen School of Medicine, University of California, Los Angeles, California 90095, USA. ³Department of Biostatistics, School of Public Health, University of California, Los Angeles, California 90095, USA. ⁴Department of Biological Chemistry, David Geffen School of Medicine, University of California, Los Angeles, California 90095, USA.

*These authors contributed equally to this work.

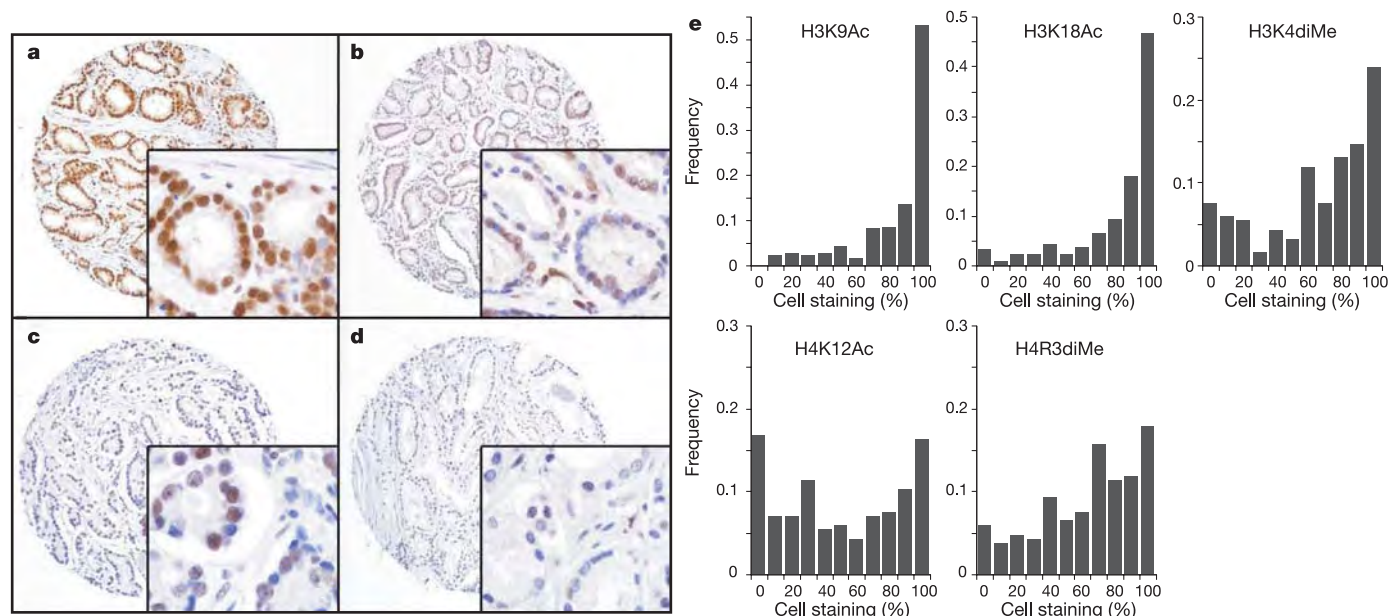


Figure 1 | Global levels of individual histone modifications are determined by immunohistochemistry. **a–d**, Characteristic nuclear staining of malignant prostate glandular cells by immunohistochemistry with antibodies against H3 K18 acetylation (**a**, **b**) or against H4 R3 dimethylation (**c**, **d**). Representative sections from group 1 patients with 95% (**a**) and 70%

(**c**) positive staining and of group 2 patients with 50% (**b**) and 25% (**d**) positive staining are shown (see Fig. 3). Original magnifications, $\times 10$; insets, $\times 40$. **e**, Distribution of staining for the five different antibodies across all 183 tissue samples. The y axis is the fraction of samples showing positive staining for the indicated percentage of cells (x axis).

H4 K12Ac ($r = 0.33$, $P = 4.8 \times 10^{-6}$) and H4 R3diMe ($r = 0.23$, $P = 2.0 \times 10^{-3}$) are positively correlated with increasing grade but H3 K9Ac ($r = 0.11$, $P = 1.4 \times 10^{-1}$) shows no significant correlation. Despite the positive correlations with grades, none of the modifications is associated individually with the risk of tumour recurrence (Supplementary Table S1). The significance of these correlations is unclear, but the higher levels of staining might be related to the increased proliferative capacity of dedifferentiated tumours, which might be associated with increased gene activity. In this regard, H3 K18Ac and H4 R3diMe are two histone modifications associated with gene activity^{17,18}.

The relationships described above are between individual modifications and grade for all patient samples. To determine whether unique patterns of histone modifications, involving combinations of the five sites, were shared between subsets of tissue samples, we

applied the random forest (RF) clustering algorithm to the data^{19–21}. RF clustering is an unsupervised classification method that, by generating an ensemble of individual tree predictors, leads to a measure of natural dissimilarity between the observations. The result of the RF clustering of all 183 samples is shown in Fig. 2a as a multidimensional scaling (MDS) plot. In MDS plots, the orientation and unit of axes are arbitrary but increasing distances between data points reflect increasing degree of dissimilarity. Each patient sample is labelled according to its Gleason score. As delineated by the blue line in Fig. 2a, two groups of patients were identified by inspection and are colour-coded as red ($n = 70$) and black ($n = 113$). In Kaplan–Meier²² survival analysis, we did not detect a statistically significant difference in the risk of tumour recurrence between these two groups (Fig. 2b).

However, further stratification of patients into high-grade (Gleason

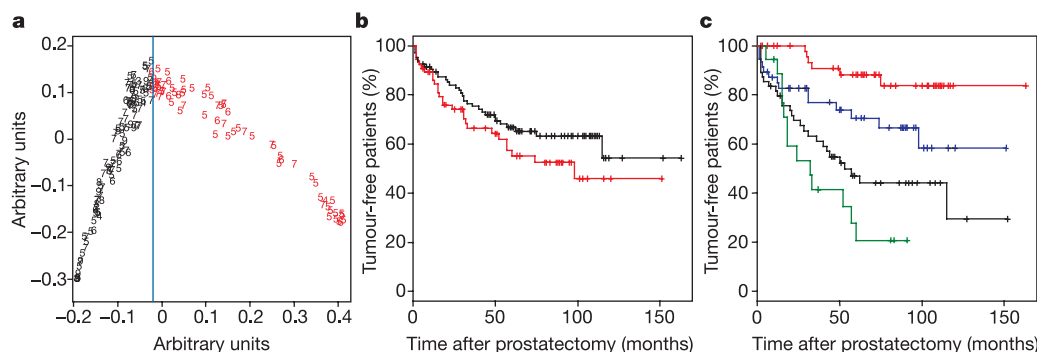


Figure 2 | Grouping of patients with similar histone modification patterns. **a**, An MDS plot is used to visualize the degree of dissimilarity (distance along the axes with arbitrary units) between all patients (with low-grade and high-grade tumours) as generated by the RF algorithm on the basis of the histone modification patterns. Patients are indicated by their Gleason score. Two groups are identified by inspection (blue line dividing

the sample into group 1 (black) and group 2 (red)). **b**, Kaplan–Meier recurrence-free plots of the two groups (black, group 1; red, group 2) identified among all patients. Log-rank $P = 0.178$. **c**, Kaplan–Meier recurrence-free survival plots based on modification pattern grouping (as in **a**) and grade stratification (blue and red lines, patients with low-grade tumours; green and black lines, patients with high-grade tumours).

score 7–10) and low-grade (Gleason score 2–6) patients revealed two subgroups within each category of patients with a considerable difference in their risk of tumour recurrence (Fig. 2c; compare red with blue lines and green with black lines). These results indicate that the differences in histone modification patterns might be a better predictor of clinical outcome when patients are first stratified broadly on the basis of grade. We therefore first stratified patients into those with a Gleason score of 7 or more ($n = 79$) and those with a Gleason score of less than 7 ($n = 104$; Supplementary Table S2), and then applied the RF clustering algorithm. In the patients with high-grade tumours, distinct groups based on histone modifications were not clearly evident (data not shown). However, among the patients with low-grade tumours we found two groups of patients by inspection (Fig. 3a). The median levels and distribution of staining for the various modifications in the two groups are shown in the box plots²³ of Fig. 3b. As expected from the clustering, the two groups show different modification patterns. For instance, the median per cent cell staining for H3K9Ac in group 1 is 90%, whereas that in group 2 is 16%. Within each group of patients, different histone modifications showed differential levels of staining. For instance, within group 1, the median percentage cell staining for H3K9Ac was 100%, whereas that for H4K12Ac was 68%. Similarly, within group 2, the median percentage cell staining for H3K18Ac was 52%, whereas that for H4K12Ac was 5%. Taken together, these observations indicate that groups of patients can be identified on the basis of similar combinations of global histone modifications.

To ascertain whether the identified groups are clinically significant, we determined the risk of tumour recurrence in each group after removal of the primary tumour (Fig. 4a). Remarkably, the patients in group 1 had a lower risk of 10-year tumour recurrence (17%) when compared with those in group 2 (42%) ($P = 0.0076$). Grade does not substitute for the histone modifications because there was no significant difference in the distribution of patients based on the Gleason score between the two groups (Fisher's exact test, $P > 0.2$; Fig. 4a). We also asked whether the modification patterns added prognostic information beyond other known prognostic factors. We found that the histone modification patterns predicted tumour recurrence independently of tumour stage, preoperative PSA, and capsule invasion (Table 1). Specific patterns of global histone modifications therefore represent independent molecular markers associated with distinct clinical outcomes.

Although all five modification sites contributed to the grouping of patients above, the groups identified can also be estimated from less information. For instance, group 1 individuals with a lower risk of tumour recurrence can be identified as those patients who are above

Table 1 | Multivariate proportional hazard analysis

Variable	Hazard ratio	95% CI	P
Tumour stage	7.54	(1.86–30.47)	0.0046
Preoperative serum PSA (ng ml ⁻¹)	1.02	(0.98–1.07)	0.3100
Capsule invasion	3.41	(1.40–8.30)	0.0070
Histone modification patterns	3.86	(1.18–12.62)	0.0250

the 60 percentile staining for H3K4diMe or above 35 percentile staining for H3K18Ac and above 35 percentile staining for H3K4diMe; those patients that do not satisfy this rule belong to group 2 (Supplementary Fig. S2). When this rule is used to predict recurrence risk, it results in only two misclassified patients (log-rank $P = 0.028$; hazard ratio = 2.8; 95% confidence interval (CI) 1.12–7.02). Thus, although additional information leads to more significant groupings, simpler rules involving a limited number of modifications can be constructed that might prove to be more practicable in clinical settings.

To validate the prognostic power of histone modifications, an additional independent set of 39 patient samples with low-grade prostate cancer (obtained from the University of Michigan Medical School) was analysed according to the above simple rule involving H3K18Ac and K4diMe staining (Supplementary Table S3). As shown in Fig. 4b, the staining distinguishes between two groups of patients with distinct risks of tumour recurrence: 4% in group A versus 31% in group B (log-rank $P = 0.016$; hazard ratio = 9.2; 95% CI 1.02–82.2). There is no significant difference in the distribution of patients based on the Gleason score between the two groups (Fisher's exact test, $P > 0.2$; Fig. 4b). The prognostic classification on the validation set therefore confirms the predictive power of histone modifications as markers of prognosis.

We have provided evidence that changes in bulk histone modifications of cancer cells are predictive of clinical outcome. The mechanistic basis of such changes are currently unclear but may be related to the altered expression and/or global activities of various histone-modifying enzymes. The variability in the levels of any one modification was not sufficient for predicting outcome. However, in combination, these changes proved to be indicative of the risk of tumour recurrence in patients with low-grade prostate cancer. Considering the substantial number of modifications on histones, it is possible that information on global patterns of other modification sites will help with the further classification of all patients, including those in the high-grade category. The utility of immuno-histochemistry, combined with the availability of an extensive set of

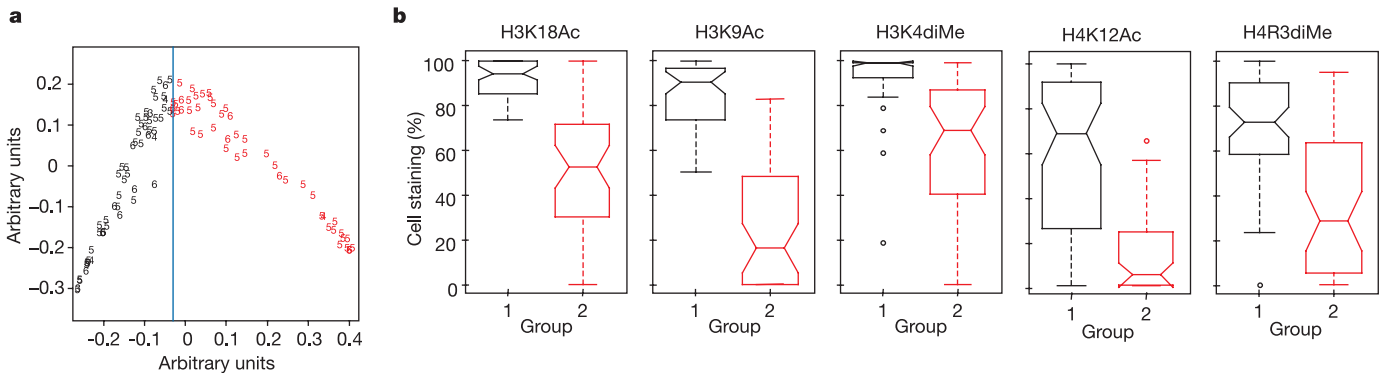


Figure 3 | Grouping of patients with low-grade tumours with similar histone modification patterns. **a**, An MDS plot showing the degree of dissimilarity between patients with low-grade tumours as generated by the RF algorithm. Patients are indicated by their Gleason score. Two groups are identified by inspection (blue line dividing the sample into group 1 (black)

and group 2 (red)). **b**, The distributions of staining for the five histone modifications in group 1 (black) and group 2 (red) patients are shown as box plots. The line in the centre of each box represents the median value of the distribution, and the upper and lower ends of the box are the upper (25th) and lower (75th) quartiles, respectively. The whiskers show the ranges.

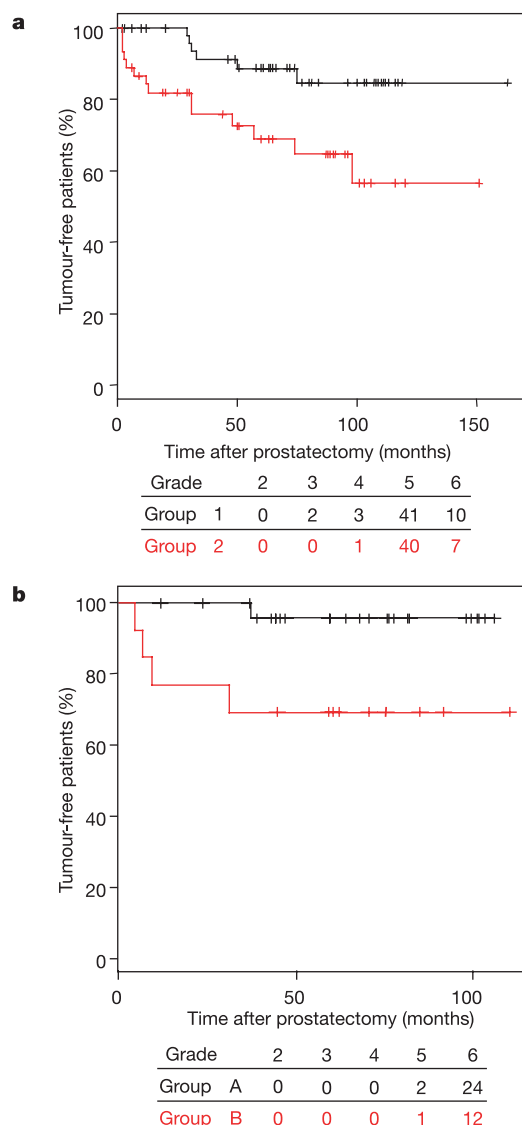


Figure 4 | Histone modification patterns predict tumour recurrence. **a**, Kaplan–Meier recurrence-free plots of the two groups (black, group 1; red, group 2) identified among the patients with low-grade tumours ($n = 104$; UCLA) based on the histone modification patterns. Log-rank $P = 0.0076$. Tabulated below is the distribution of the patients in each group according to their grade (Gleason score). **b**, Kaplan–Meier recurrence-free plots of the two groups (black, group A; red, group B) identified among the patients with low-grade tumours of the validation data set ($n = 39$; University of Michigan). The two groups are identified on the basis of the ‘simple rule’ involving only H3 K18Ac and H3 K4diMe modifications. Log-rank $P = 0.016$. The distribution of the patients in each group according to their grade is tabulated below.

antibodies to probe histone modifications, should facilitate the application of our approach to other tumours. It is conceivable that patterns of histone modifications including those reported here might also serve as prognostic or even diagnostic markers in other types of cancer.

Note added in proof: While this paper was being peer-reviewed, it was reported that acetylation of H4 K16 and trimethylation of H4 K20 are reduced at repetitive DNA sequences in multiple cancer types²⁴.

METHODS

Prostate TMA. A prostate TMA was constructed with formalin-fixed, paraffin-embedded prostate tissue samples as described previously¹⁶. At least three replicate tumour samples were taken from donor tissue blocks in a highly

representative fashion. Twenty patients treated with neoadjuvant hormones were excluded from the study. In total, 183 cases were informative for all five histone markers; 171 of those were supported by complete recurrence data. A retrospective analysis for outcome assessment was based on detailed anonymized clinico-pathological information linked to the TMA specimens. Recurrence, defined as a postoperative serum PSA of 0.2 ng ml^{-1} or more, was seen in 61 (34%) of all study patients and in 20 (19%) of patients with low-grade tumours. The median total follow-up, defined as the time to recurrence or to last contact in non-recurring patients, was 60.0 months (range 2–163) for patients with low-grade tumours. The median follow-up time within the recurring and non-recurring patient groups was 30.5 months (2.0–98.0) and 65.5 months (range 2.0–163.0), respectively, in patients with low-grade tumours. The validation data set was generated from prostate TMAs that were purchased from the University of Michigan Medical School (Supplementary Table S3).

Immunohistochemistry. The antibodies were first tested and optimized on whole-tissue sections and test arrays. Once an appropriate dilution had been determined, a set of three slides containing all patient samples were stained for each antibody, using standard two-step indirect immunohistochemistry. Tissue array sections were cut with of a sectioning aid (Instrumedics) immediately before being stained. After deparaffinization in xylenes, the sections were rehydrated in graded alcohols. Endogenous peroxidase was quenched with 3% hydrogen peroxide in methanol at room temperature (25°C). The sections were placed in a 95°C solution of 0.01 M sodium citrate buffer pH 6.0 for antigen retrieval. Normal goat serum (5%) was next applied for 30 min to block non-specific protein-binding sites. Primary rabbit anti-histone polyclonal antibodies were applied for 30 min at room temperature at the following dilutions: H3 K18Ac at 1:200, H3 K9Ac at 1:800, H4 K12Ac at 1:100, H3 K4diMe (Abcam) at 1:800, and H4 R3diMe (Upstate) at 1:25. Detection was accomplished with the Dako Envision System, followed by chromogen detection with diaminobenzidine (DAB). The sections were counterstained with Harris’s haematoxylin, followed by dehydration and mounting. Negative controls were identical array sections stained in the absence of the primary antibody. Semiquantitative assessment of antibody staining on the TMAs was performed by H.Y. and D.B.S., who were blinded to all clinico-pathological variables. The frequency of nuclear positive target cells (range 0–100%) in prostatic glandular epithelium was scored for each TMA spot.

Unsupervised clustering algorithm. To facilitate unsupervised learning, an intrinsic dissimilarity measure between the patients was constructed with an RF analysis of the histone markers. A technical description of the RF clustering algorithm is given in Supplementary Methods and <http://www.genetics.ucla.edu/labs/horvath/RFclustering/RFclustering.htm>. The RF clustering algorithm was shown recently to be particularly suitable for TMA data for the following reasons²¹. First, the clustering results do not change when one or more covariates are monotonically transformed, because the dissimilarity depends only on the feature ranks, obviating the need for symmetrizing skewed covariate distributions. Second, the RF dissimilarity weights the contributions of each covariate on the dissimilarity in a natural way: the more related the covariate is to other covariates the more it will affect the definition of the RF dissimilarity. Third, the RF dissimilarity does not require the user to specify threshold values for dichotomizing tumour expressions. External threshold values for dichotomizing expressions in unsupervised analyses may reduce the information content or even bias the results. We also compared the RF clustering approach with the standard euclidean distance-based approach. Although there is good overlap between the two algorithms, we find that the RF clustering method works better for these data (Supplementary Methods). To reveal the clustering, we used classical MDS, which takes as input the RF dissimilarity between the samples and returns a set of points in a two-dimensional space such that the distances between the points are approximately equivalent to the original distances.

Statistical analysis. To test whether variables differed across groups, we used the Kruskal–Wallis test. To visualize the survival distributions, we used Kaplan–Meier plots. The Cox proportional hazards model was used to test the statistical independence and significance of predictors. The proportional hazard assumption was tested by using scaled Schoenfeld residuals. Log-rank tests were used to test the difference between survival distributions. All P values were two-sided, and $P < 0.05$ was considered significant. All statistical analyses were performed with the freely available software R (<http://www.R-project.org/>). R code that implements RF clustering is available from the authors on request.

Further details. A more detailed description of the methods used is given in the Supplementary Methods.

Received 21 February; accepted 25 April 2005.

1. Jacobson, S. & Pillus, L. Modifying chromatin and concepts of cancer. *Curr. Opin. Genet. Dev.* 9, 175–184 (1999).

2. Vogelauer, M., Wu, J., Suka, N. & Grunstein, M. Global histone acetylation and deacetylation in yeast. *Nature* **408**, 495–498 (2000).
3. Jemal, A. *et al.* Cancer statistics, 2003. *CA Cancer J. Clinicians* **53**, 5–26 (2003).
4. Gleason, D. F. Classification of prostatic carcinomas. *Cancer Chemother. Rep.* **50**, 125–128 (1966).
5. Bunting, P. S. Screening for prostate cancer with prostate-specific antigen: beware the biases. *Clin. Chim. Acta* **315**, 71–97 (2002).
6. Han, M., Partin, A. W., Piantadosi, S., Epstein, J. I. & Walsh, P. C. Era specific biochemical recurrence-free survival following radical prostatectomy for clinically localized prostate cancer. *J. Urol.* **166**, 416–419 (2001).
7. Farkas, A., Schneider, D., Perrotti, M., Cummings, K. B. & Ward, W. S. National trends in the epidemiology of prostate cancer, 1973 to 1994: evidence for the effectiveness of prostate-specific antigen screening. *Urology* **52**, 444–448 (1998).
8. Giles, R. H., Peters, D. J. & Breuning, M. H. Conjunction dysfunction: CBP/p300 in human disease. *Trends Genet.* **14**, 178–183 (1998).
9. Gayther, S. A. *et al.* Mutations truncating the EP300 acetylase in human cancers. *Nature Genet.* **24**, 300–303 (2000).
10. Muraoka, M. *et al.* p300 gene alterations in colorectal and gastric carcinomas. *Oncogene* **12**, 1565–1569 (1996).
11. Debes, J. D. *et al.* p300 in prostate cancer proliferation and progression. *Cancer Res.* **63**, 7638–7640 (2003).
12. Reid, J. L., Iyer, V. R., Brown, P. O. & Struhl, K. Coordinate regulation of yeast ribosomal protein genes is associated with targeted recruitment of Esa1 histone acetylase. *Mol. Cell* **6**, 1297–1307 (2000).
13. Krebs, J. E., Fry, C. J., Samuels, M. L. & Peterson, C. L. Global role for chromatin remodeling enzymes in mitotic gene expression. *Cell* **102**, 587–598 (2000).
14. Peterson, C. L. & Laniel, M. A. Histones and histone modifications. *Curr. Biol.* **14**, R546–R551 (2004).
15. Suka, N., Suka, Y., Carmen, A. A., Wu, J. & Grunstein, M. Highly specific antibodies determine histone acetylation site usage in yeast heterochromatin and euchromatin. *Mol. Cell* **8**, 473–479 (2001).
16. Kononen, J. *et al.* Tissue microarrays for high-throughput molecular profiling of tumour specimens. *Nature Med.* **4**, 844–847 (1998).
17. Kurdastani, S. K., Tavazoie, S. & Grunstein, M. Mapping global histone acetylation patterns to gene expression. *Cell* **117**, 721–733 (2004).
18. Rezai-Zadeh, N. *et al.* Targeted recruitment of a histone H4-specific methyltransferase by the transcription factor YY1. *Genes Dev.* **17**, 1019–1029 (2003).
19. Breiman, L. *Classification and Regression Trees* (Wadsworth International, Belmont, California, 1984).
20. Breiman, L. Random forests. *Mach. Learn.* **45**, 5–32 (2001).
21. Shi, T., Seligson, D., Beldegrun, A. S., Palotie, A. & Horvath, S. Tumor classification by tissue microarray profiling: random forest clustering applied to renal cell carcinoma. *Mod. Pathol.* **18**, 547–557 (2005).
22. Kaplan, E. & Meier, P. Nonparametric estimation from incomplete observations. *J. Am. Stat. Assoc.* **53**, 457–481 (1958).
23. Cleveland, W. S. *Visualizing Data* (Hobart Press, Murray Hill, New Jersey, 1993).
24. Fraga, M. F. *et al.* Loss of acetylation at Lys16 and trimethylation at Lys20 of histone H4 is a common hallmark of human cancer. *Nature Genet.* **37**, 391–400 (2005).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank M. Vogelauer for suggestions on the manuscript, and V. Minin for help with the statistical analyses. T.S. was a doctoral trainee supported by the UCLA Integrative Graduate Education and Research Traineeship (IGERT) Bioinformatics Program funded by the NSF Division of Graduate Education (DGE). This work was funded partly by a National Cancer Institute (NCI) grant through the Jonsson Comprehensive Cancer Center to D.B.S. and a Howard Hughes Medical Institute Fellowship and a UCLA Specialized Program Of Research Excellence (SPORE) in Prostate Cancer Career Development grant to S.K.K.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to S.K.K. (skurdastani@mednet.ucla.edu).

Dynamics of chronic myeloid leukaemia

Franziska Michor¹, Timothy P. Hughes², Yoh Iwasa³, Susan Branford², Neil P. Shah⁴, Charles L. Sawyers^{4,5} & Martin A. Nowak¹

The clinical success of the ABL tyrosine kinase inhibitor imatinib in chronic myeloid leukaemia (CML) serves as a model for molecularly targeted therapy of cancer^{1–4}, but at least two critical questions remain. Can imatinib eradicate leukaemic stem cells? What are the dynamics of relapse due to imatinib resistance, which is caused by mutations in the ABL kinase domain? The precise understanding of how imatinib exerts its therapeutic effect in CML and the ability to measure disease burden by quantitative polymerase chain reaction provide an opportunity to develop a mathematical approach. We find that a four-compartment model, based on the known biology of haematopoietic differentiation⁵, can explain the kinetics of the molecular response to imatinib in a 169-patient data set. Successful therapy leads to a biphasic exponential decline of leukaemic cells. The first slope of 0.05 per day represents the turnover rate of differentiated leukaemic cells, while the second slope of 0.008 per day represents the turnover rate of leukaemic progenitors. The model suggests that imatinib is a potent inhibitor of the production of differentiated leukaemic cells, but does not deplete leukaemic stem cells. We calculate the probability of developing imatinib resistance mutations and estimate the time until detection of resistance. Our model provides the first quantitative insights into the *in vivo* kinetics of a human cancer.

Chronic myeloid leukaemia (CML) represents the first human cancer in which molecularly targeted therapy leads to a dramatic clinical response^{1–4}. In most patients, however, the rapid decline of the leukaemic cell burden induced by the ABL tyrosine kinase inhibitor imatinib fails to eliminate residual disease. Bone marrow studies have shown that the residual cells are part of the leukaemic stem cell compartment^{6,7}. This observation raises the question of whether imatinib is capable of impairing the proliferation of leukaemic stem cells. Moreover, a substantial fraction of patients develops acquired resistance to imatinib. Mutations in the ABL kinase domain are the main mechanism of resistance and account for 70–80% of cases with treatment failure^{8–13}. Sometimes, resistance mutations are present in leukaemic cells prior to imatinib therapy^{13–15}.

We design a mathematical model which describes four layers of the differentiation hierarchy of the haematopoietic system (see Methods and Supplementary Information). Stem cells give rise to progenitors, which produce differentiated cells, which produce terminally differentiated cells. This hierarchy applies both to normal and leukaemic cells. Therefore, the leukaemic cell population consists of leukaemic stem cells and three types of leukaemic differentiated cells; only leukaemic stem cells have an indefinite potential for self-renewal. The *BCR-ABL* oncogene is present in all leukaemic cells. It leads to a slow clonal expansion of leukaemic stem cells and accelerates the rate at which these cells produce leukaemic progenitors and differentiated cells.

We analysed 169 CML patients. The levels of *BCR-ABL* transcripts in the blood of the patients is measured by a quantitative real-time PCR (RQ-PCR) assay^{10,16}. *BCR* is used as the control gene and *BCR-ABL* values are expressed as a percentage of the *BCR* transcript levels to compensate for variations in the RNA quality and efficiency of reverse transcription. Because the blood predominantly contains terminally differentiated cells, the obtained values give an estimate of the fraction of terminally differentiated leukaemic cells.

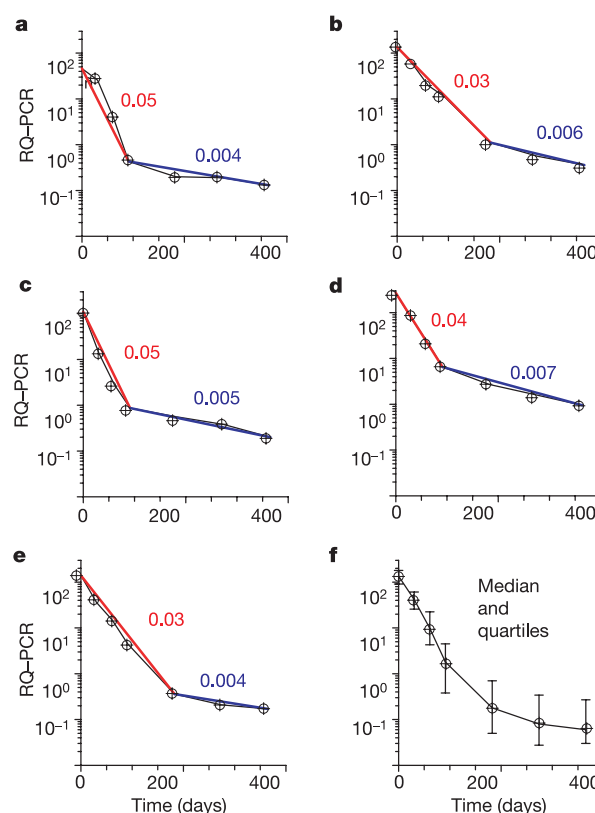


Figure 1 | Imatinib leads to a biphasic decline of leukaemic cells. a–e, The levels of *BCR-ABL* transcripts in the blood of five patients are shown during 12 months of therapy starting at day 0. In these patients, the first slope ranges from 0.03 to 0.05 per day and the second slope from 0.004 to 0.007 per day. The first slope represents the death rate of leukaemic differentiated cells and the second slope the death rate of leukaemic progenitors during imatinib therapy. Panel f shows the median with quartiles taken over all patients who do not have a rise in the leukaemic cell burden during the first 12 months of therapy. The circle represents the 50 percentile, and the bars the 25 and 75 percentiles.

¹Program for Evolutionary Dynamics, Department of Organismic and Evolutionary Biology, Department of Mathematics, Harvard University, Cambridge, Massachusetts 02138, USA. ²Institute of Medical and Veterinary Science, Adelaide, Australia. ³Department of Biology, Kyushu University, Fukuoka 812-8581, Japan. ⁴Department of Medicine, and ⁵Howard Hughes Medical Institute, Molecular Biology Institute, Department of Urology, Department of Medical and Molecular Pharmacology, Jonsson Comprehensive Cancer Center, David Geffen School of Medicine, University of California, Los Angeles, California 90095, USA.

Successful therapy leads to a biphasic exponential decline of leukaemic cells (Fig. 1). The first slope is determined by calculating the exponential decline between 0 and 3 months; a mean value of 0.05 ± 0.02 per day is obtained, which corresponds to a decline of 5% per day. The second slope is determined by calculating the exponential decline between 6 and 12 months; a mean value of 0.008 ± 0.004 per day is obtained, which corresponds to a decline of 0.8% per day. This analysis considers only patients who do not have any rise in the leukaemic cell burden during the first 12 months of therapy, in order to exclude the effect of acquired resistance.

Fitting our model to the data, we conclude that the first slope represents the turnover rate of differentiated leukaemic cells. Therefore, these cells have an average lifespan of $1/0.05 = 20$ days. Upon reaching a steady state with the leukaemic progenitors, the number of differentiated leukaemic cells decreases at the same rate as the number of leukaemic progenitors. The second slope represents the turnover rate of leukaemic progenitors. Hence, these cells have an average lifespan of $1/0.008 = 125$ days. Both estimates denote average lifespans during imatinib therapy. Any process that removes cells from the corresponding subpopulation (including further differentiation) contributes to the average lifespan.

The first slope leads to an approximately 1,000-fold decline in the leukaemic cell burden. Therefore, imatinib reduces the rate at which leukaemic differentiated cells arise from leukaemic progenitors about 1,000-fold. This effect is as if imatinib prevented about ten rounds of cell division of leukaemic cells ($2^{10} = 1,024$), either by increasing their death rate or by reducing their division rate.

Some patients ceased imatinib therapy because of complications or side effects (Fig. 2). Even if imatinib had been administered for many months (up to three years), the numbers of BCR-ABL transcripts in those patients rose within three months after discontinuation of therapy to levels at pre-treatment baseline or above. Other studies of patients who ceased imatinib led to similar findings^{17,18}. We conclude that long-term imatinib treatment does not deplete the cell population that drives this disease. This conclusion is consistent with the hypothesis that leukaemic stem cells are

insensitive to chemotherapy^{19–21}. The average rate of the exponential increase after stopping treatment is 0.09 ± 0.05 per day, corresponding to a doubling time of 8 days. This timescale indicates the rate at which terminally differentiated leukaemic cells arise from leukaemic stem cells in the absence of imatinib.

A comparison between model and data suggests the following two concepts. (1) Imatinib treatment leads to a competitive disadvantage of leukaemic progenitors and leukaemic differentiated cells; their production rates are dramatically reduced. The consequence is a biphasic decline of the abundance of the BCR-ABL transcript in response to therapy. (2) Leukaemic stem cells, however, are not depleted during imatinib treatment. The total leukaemic cell burden rapidly returns to the baseline value (or beyond) when imatinib is discontinued.

Acquired imatinib resistance usually develops owing to mutations in the ABL kinase domain^{8–13}. Resistant leukaemic cells emerge after an initially successful response to imatinib therapy and lead to a relapse of disease (Fig. 3). The average slope was determined by calculating the exponential increase after the first appearance of resistance mutations in 30 patients; a value of 0.02 ± 0.01 per day was obtained. Of those patients who start imatinib in the early chronic, late chronic, and accelerated phase of CML, respectively, 12%, 32%, and 62% develop detectable resistance mutations within two years of treatment²².

Our basic model can be extended to include the stochastic evolution of resistance. Consider an exponentially growing leukaemia continually producing resistant cells at rate u per cell division. The probability of having resistance mutations once leukaemic stem cells have reached a certain abundance, y_0 , is given by $P = 1 - \exp(-uy_0\sigma)$ where $\sigma = (1 + s)\log(1 + 1/s)$. The parameter s denotes the excess reproductive ratio of the exponentially growing leukaemia, which is the relative difference between birth and death rates. This calculation assumes that resistance mutations are neutral prior to therapy.

If the point mutation rate is about 10^{-8} per base per cell division²³, then resistance due to any one of about forty known mutations^{7–14}

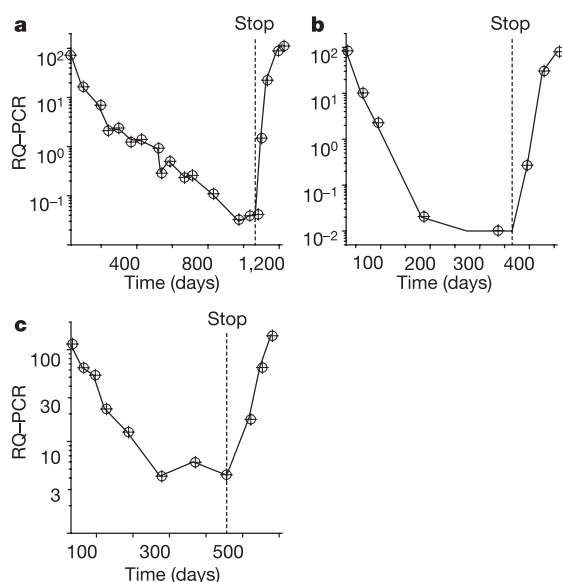


Figure 2 | Discontinuation of imatinib therapy in three patients after 1–3 years led to a rapid increase of leukaemic cells to levels at or beyond pre-treatment baseline. We conclude that leukaemic stem cells, which drive CML disease, are not depleted by imatinib therapy. The rapid upslope of 0.09 ± 0.05 per day corresponds to a doubling time of roughly 8 days, which characterizes the rate at which differentiated leukaemic cells are regenerated from leukaemic stem cells. Each of panels **a**, **b** and **c** corresponds to one of the three patients.

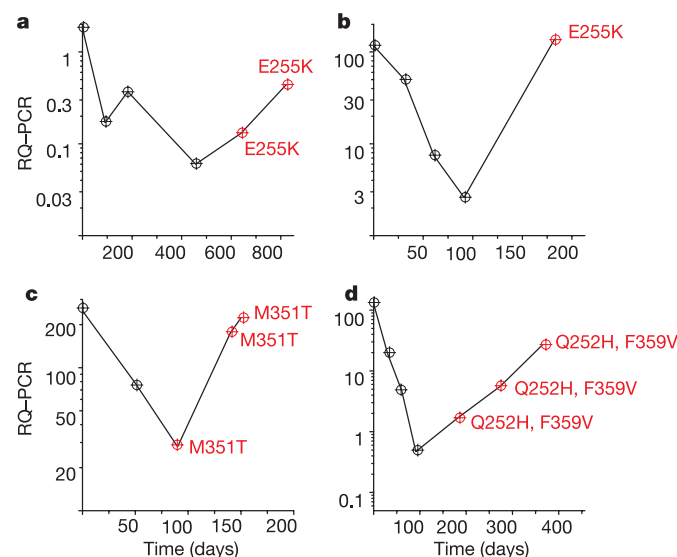


Figure 3 | About 40 different point mutations in BCR-ABL have been identified that confer various degrees of resistance to imatinib therapy. **a–d**, We show evolutionary dynamics of resistance in four patients. The labels denote the individual mutations that are detected at various time points. Resistance mutations lead to a relapse of leukaemic cells. The upslope ranges from 0.003 to 0.06 per day in 30 patients, with an average of 0.02 ± 0.01 per day. The characteristic timescale for the rise of resistance is given by the rate at which resistant leukaemic stem cells expand during therapy.

arises at rate $u = 4 \times 10^{-7}$ per cell division. The abundance of leukaemic stem cells at diagnosis in early chronic phase^{24,25} is estimated to be about $y_0 = 2.5 \times 10^5$. For these values and $s = 1$, we calculate that about 13% of patients harbour resistance mutations at the time of diagnosis. If imatinib therapy commences at a later stage of the disease, when the leukaemic stem cell population has expanded to 10^7 cells, for example, then 100% of those patients have some resistant leukaemic stem cells. Therefore, the higher incidence of resistance in patients who start imatinib therapy in a later phase of disease can be explained by an increased leukaemic stem cell burden.

We expect that the frequency of resistance mutations at the start of imatinib therapy is below detection limit in most patients. We can use the deterministic model to calculate the time until detection of resistance mutations and treatment failure (see Supplementary Information). In our model, the characteristic timescale is given by the rate at which the resistant leukaemic stem cells are expanding during therapy. Therefore, a faster-growing leukaemia leads to an earlier emergence of resistance, while a slower-growing leukaemia might allow many years of successful therapy, even if resistant cells are present at low frequencies. The reason for this unusual behaviour is that leukaemic stem cells, with or without resistance mutations, continue to expand during treatment. Imatinib acts by reducing the abundance of leukaemic progenitors, differentiated and terminally differentiated cells without resistance mutations.

There is evidence that some resistance mutations confer a growth advantage even in the absence of imatinib^{8,13}. These 'advantageous mutations' have a higher probability of being present in patients prior to treatment and can lead to a more rapid expansion during therapy.

Figure 4 summarizes the dynamical features of our mathematical model. Normal haematopoietic cells are in a steady state. Leukaemic stem cells expand exponentially at a slow rate. Imatinib reduces the rate at which leukaemic stem cells produce progenitors. Hence, the abundance of leukaemic progenitors declines once treatment is started. Similarly imatinib reduces the rate at which leukaemic progenitors produce differentiated cells. The abundance of differentiated leukaemic cells shows a biphasic decline. The first slope is determined by the average lifespan of leukaemic differentiated cells (20 days), while the second slope reflects the longer lifespan of leukaemic progenitors (125 days). We assume that imatinib does not affect the rate at which leukaemic differentiated cells produce terminally differentiated cells. This cell population has a fast turnover rate—on a timescale of one day—and simply tracks the biphasic decline of leukaemic differentiated cells. If imatinib is discontinued, there is a rapid resurgence of the leukaemic load because the cell population which drives the disease, the leukaemic stem cells, was not depleted during therapy. Resistant leukaemic stem cells might expand faster than leukaemic stem cells during therapy for two reasons: (1) either they have an inherent selective advantage; or (2) imatinib somewhat reduces the growth rate of leukaemic stem cells without depleting them. In any case, resistant leukaemic stem cells continue to produce large amounts of progenitors and differentiated cells during therapy. The total leukaemic cell burden declines initially, but rises again once resistant differentiated cells become abundant. In this case, the time to treatment failure is determined by the rate of expansion of resistant leukaemic stem cells.

Finding a cellular mechanism for drug resistance of leukaemic stem cells is a very important goal for future experimental research. Imatinib is a substrate for the multidrug resistance protein MDR p-glycoprotein and will therefore be excluded from cells that express significant MDR levels²⁶. Stem cells naturally express higher levels of MDR²⁷. It is unknown whether this is the actual mechanism for sparing leukaemic stem cells, because it has not been possible neither to measure accurately the imatinib concentration in leukaemic stem cells nor to measure BCR-ABL inhibition selectively in this compartment. Another possibility is that leukaemic stem cells are less dependent on BCR-ABL for growth and survival than are committed

progenitors, and therefore BCR-ABL inhibition does not eliminate leukaemic stem cells. Indeed, there is evidence that BCR-ABL messenger RNA, but not protein, can be detected in a progenitor cell population²⁸.

In cell-culture systems using CML cell lines or murine haematopoietic cells transformed by BCR-ABL, imatinib leads to rapid inhibition of ABL kinase activity and subsequent induction of apoptotic cell death (on a timescale of hours). This very fast response to imatinib is in stark contrast to the 20 and 125 day half-lives we find in our *in vivo* data. This discrepancy might be explained by the fact that *in vitro* model systems are derived from (or resemble) blast crisis cells, whereas our data refer to chronic-phase CML. Since there are no *in vitro* models of chronic-phase CML, we cannot make a direct comparison. Curiously, the rate of clearance of leukaemic blasts from the blood of CML blast crisis patients is rapid (3–7 days) and may reflect increased dependence of blasts on the action of the BCR-ABL oncogene. In contrast, leukaemic cells from chronic-phase patients are not as dependent on the BCR-ABL signal and consequently do not undergo rapid apoptosis. Our analysis suggests that in chronic-phase

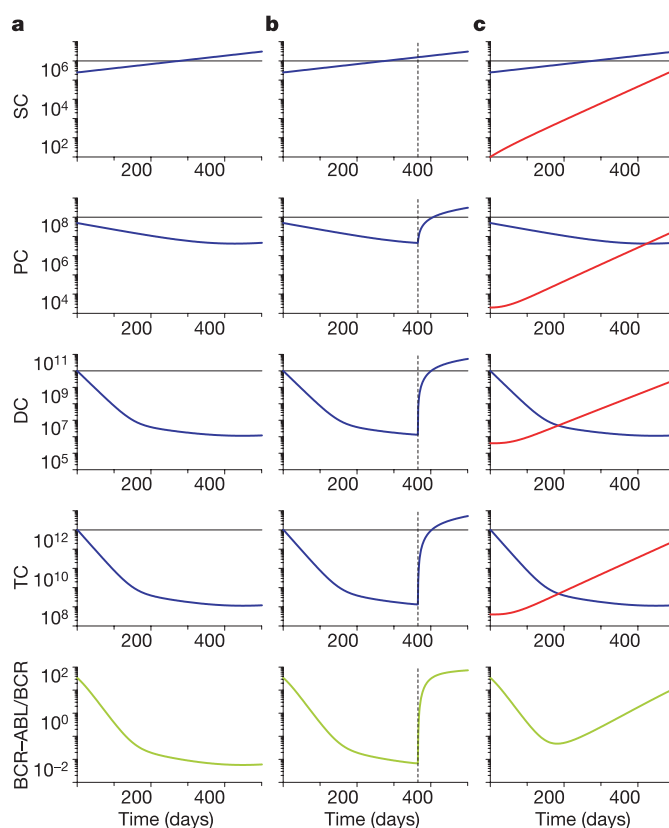


Figure 4 | Model dynamics of different treatment responses to imatinib.

a, Without resistance mutations; **b**, when therapy is stopped after one year; and **c**, with resistance mutations. The rows show stem cells (SC), progenitor cells (PC), differentiated cells (DC), and terminally differentiated cells (TC); wild type cells are black, leukaemic cells blue, and resistant leukaemic cells red. The bottom row shows the ratio of BCR-ABL over BCR in % (green). **a**, Imatinib therapy is started at day 0. Treatment leads to a biphasic decline of the BCR-ABL/BCR ratio. Leukaemic stem cells continue to expand at a slow rate. **b**, Discontinuation of treatment (broken line) leads to a rapid rise of leukaemic cells to levels above the pre-treatment baseline, because leukaemic stem cells have not been depleted during therapy. **c**, Emergence of resistance mutations leads to an increase of the BCR-ABL/BCR ratio at a rate which is determined by the rise of resistant leukaemic stem cells. Parameter values are $d_0 = 0.003$, $d_1 = 0.008$, $d_2 = 0.05$, $d_3 = 1$, $a_x = 0.8$, $b_x = 5$, $c_x = 100$, $r_y = 0.008$, $a_y = 2a_x$, $b_y = 2b_x$, $c_y = c_x$, $r_z = 0.023$. During therapy we have $a'_y = a_y/100$, $b'_y = b_y/750$, $c'_y = c_y$, $a_z = a'_z = a_y$, $b_z = b'_z = b_y$, and $c_z = c'_z = c_y$. In **c**, we have $z_0(0) = 10$ and $u = 4 \times 10^{-8}$.

CML the mode of action of imatinib is to reduce the rate at which differentiated leukaemic cells are produced from progenitors and at which progenitors are produced from leukaemic stem cells. Hence, the *in vivo* action does not rely on massive and sudden apoptosis, but on a decline in the net proliferation potential of leukaemic cells.

We conclude that the comparison between model and data provides quantitative insights into the *in vivo* kinetics of CML. We obtain numerical estimates for the turnover rates of leukaemic progenitors and differentiated cells. Imatinib dramatically reduces the rate at which those cells are being produced from leukaemic stem cells, but it does not lead to an observable decline of leukaemic stem cells. The probability of harbouring resistance mutations increases with disease progression as a consequence of an increased leukaemic stem cell abundance. The characteristic time to treatment failure caused by acquired resistance is given by the growth rate of the leukaemic stem cells. Thus, multiple drug therapy²⁹ is especially important for patients who are diagnosed with advanced and rapidly growing disease.

METHODS

The basic mathematical model. The abundances of normal haematopoietic stem cells, progenitors, differentiated cells, and terminally differentiated cells are denoted x_0 , x_1 , x_2 , and x_3 . Their respective leukaemic abundances are given by y_0 , y_1 , y_2 and y_3 for cells without resistance mutations and by z_0 , z_1 , z_2 and z_3 for cells with resistance mutations. Stem cells produce progenitors, which produce differentiated cells, which produce terminally differentiated cells. The rate constants are given by a , b and c with appropriate indices distinguishing between healthy, leukaemic and resistant cell lineages. The death rates of stem cells, progenitors, differentiated and terminally differentiated cells are denoted by d_0 , d_1 , d_2 and d_3 . Homeostasis of normal stem cells is achieved by an appropriate declining function, λ . Leukaemic stem cells divide at rate r_y , and resistant stem cells divide at rate r_z . The basic mathematical model is given by:

$$\begin{aligned}\dot{x}_0 &= [\lambda(x_0) - d_0]x_0 & \dot{y}_0 &= [r_y(1 - u) - d_0]y_0 & \dot{z}_0 &= (r_z - d_0)z_0 + r_y y_0 u \\ \dot{x}_1 &= a_x x_0 - d_1 x_1 & \dot{y}_1 &= a_y y_0 - d_1 y_1 & \dot{z}_1 &= a_z z_0 - d_1 z_1 \\ \dot{x}_2 &= b_x x_1 - d_2 x_2 & \dot{y}_2 &= b_y y_1 - d_2 y_2 & \dot{z}_2 &= b_z z_1 - d_2 z_2 \\ \dot{x}_3 &= c_x x_2 - d_3 x_3 & \dot{y}_3 &= c_y y_2 - d_3 y_3 & \dot{z}_3 &= c_z z_2 - d_3 z_3\end{aligned}$$

For further discussion and analysis of the model, see Supplementary Information.

Received 29 November 2004; accepted 21 April 2005.

1. Druker, B. J. *et al.* Effects of a selective inhibitor of the Abl tyrosine kinase on the growth of Bcr-Abl positive cells. *Nature Med.* **2**, 561–566 (1996).
2. Sawyers, C. L. *et al.* Imatinib induces hematologic and cytogenetic responses in patients with chronic myelogenous leukemia in myeloid blast crisis: results of a phase II study. *Blood* **99**, 3530–3539 (2002).
3. Holtz, M. S. *et al.* Imatinib mesylate (STI571) inhibits growth of primitive malignant progenitors in chronic myelogenous leukemia through reversal of abnormally increased proliferation. *Blood* **99**, 3792–3800 (2002).
4. Hughes, T. P. *et al.* Frequency of major molecular responses to imatinib or interferon alpha plus cytarabine in newly diagnosed chronic myeloid leukemia. *N. Engl. J. Med.* **349**, 1423–1432 (2003).
5. Spangrude, G. J., Heimfeld, S. & Weissman, I. L. Purification and characterization of mouse hematopoietic stem cells. *Science* **241**, 58–62 (1988).
6. Bhatia, R. *et al.* Persistence of malignant hematopoietic progenitors in chronic myelogenous leukemia patients in complete cytogenetic remission following imatinib mesylate treatment. *Blood* **101**, 4701–4707 (2003).
7. Chu, S. *et al.* Detection of BCR-ABL kinase mutations in CD34+ cells from chronic myelogenous leukemia patients in complete cytogenetic remission on imatinib mesylate treatment. *Blood* **105**, 2093–2098 (2005).
8. Gorre, M. E. *et al.* Clinical resistance to STI-571 cancer therapy caused by BCR-ABL gene mutation or amplification. *Science* **293**, 876–880 (2001).
9. Branford, S. *et al.* High frequency of point mutations clustered within the adenosine triphosphate-binding region of BCR/ABL in patients with chronic myeloid leukemia or Ph-positive acute lymphoblastic leukemia who develop imatinib (STI571) resistance. *Blood* **99**, 3472–3475 (2002).
10. Branford, S. *et al.* Detection of BCR-ABL mutations in patients with CML treated with imatinib is virtually always accompanied by clinical resistance, and mutations in the ATP phosphate-binding loop (P-loop) are associated with a poor prognosis. *Blood* **102**, 276–283 (2003).
11. Branford, S. *et al.* Real-time quantitative PCR analysis can be used as a primary screen to identify imatinib-treated patients with CML who have BCR-ABL kinase domain mutations. *Blood* **104**, 2926–2932 (2004).
12. Hochhaus, A. *et al.* Molecular and chromosomal mechanisms of resistance to imatinib (STI571) therapy. *Leukemia* **16**, 2190–2196 (2002).
13. Shah, N. P. *et al.* Multiple BCR-ABL kinase domain mutations confer polyclonal resistance to the tyrosine kinase inhibitor imatinib (STI571) in chronic phase and blast crisis chronic myeloid leukemia. *Cancer Cell* **2**, 117–125 (2002).
14. Roche-Lestienne, C. *et al.* Several types of mutations of the Abl gene can be found in chronic myeloid leukemia patients resistant to STI571, and they can pre-exist to the onset of treatment. *Blood* **100**, 1014–1018 (2002).
15. Roche-Lestienne, C., Lai, J. L., Darre, S., Facon, T. & Preudhomme, C. A mutation conferring resistance to imatinib at the time of diagnosis of chronic myelogenous leukemia. *N. Engl. J. Med.* **348**, 2265–2266 (2003).
16. Branford, S., Hughes, T. P. & Rudzki, Z. Monitoring chronic myeloid leukaemia therapy by real-time quantitative PCR in blood is a reliable alternative to bone marrow cytogenetics. *Br. J. Haematol.* **107**, 587–599 (1999).
17. Cortes, J., O'Brien, S. & Kantarjian, H. Discontinuation of imatinib therapy after achieving a molecular response. *Blood* **104**, 2204–2205 (2004).
18. Higashi, T. *et al.* Imatinib mesylate-sensitive blast crisis immediately after discontinuation of imatinib mesylate therapy in chronic myelogenous leukemia: report of two cases. *Am. J. Hematol.* **76**, 275–278 (2004).
19. Graham, S. M. *et al.* Primitive, quiescent, Philadelphia-positive stem cells from patients with chronic myeloid leukemia are insensitive to STI571 *in vitro*. *Blood* **99**, 319–325 (2002).
20. Chaudhary, P. M. & Roninson, I. B. Expression and activity of P-glycoprotein, a multidrug efflux pump, in human hematopoietic stem cells. *Cell* **66**, 85–94 (1991).
21. Mahon, F. X. *et al.* MDR1 gene overexpression confers resistance to imatinib mesylate in leukemia cell line models. *Blood* **101**, 2368–2373 (2003).
22. Branford, S. *et al.* The incidence of BCR-ABL kinase mutations in chronic myeloid leukemia patients is as high in the second year of imatinib therapy as the first but survival after mutation detection is significantly longer for patients with mutations detected in the second year of therapy. *Blood* **102**, 414 (2003).
23. Kunkel, T. A. & Bebenek, K. DNA replication fidelity. *Annu. Rev. Biochem.* **69**, 497–529 (2000).
24. Holyoake, T. L., Jiang, X., Drummond, M. W., Eaves, A. C. & Eaves, C. J. Elucidating critical mechanisms of deregulated stem cell turnover in the chronic phase of chronic myeloid leukemia. *Leukemia* **16**, 549–558 (2002).
25. Holyoake, T., Jiang, X., Eaves, C. & Eaves, A. Isolation of a highly quiescent subpopulation of primitive leukemic cells in chronic myeloid leukemia. *Blood* **94**, 2056–2064 (1999).
26. Mahon, F. X. *et al.* MDR1 gene overexpression confers resistance to imatinib mesylate in leukemia cell line models. *Blood* **101**, 2368–2373 (2003).
27. Chaudhary, P. M. & Roninson, I. B. Expression and activity of P-glycoprotein, a multidrug efflux pump, in human hematopoietic stem cells. *Cell* **66**, 85 (1991).
28. Bedi, A. *et al.* BCR-ABL gene rearrangement and expression of primitive hematopoietic progenitors in chronic myeloid leukemia. *Blood* **81**, 2898–28902 (1993).
29. Shah, N. P. *et al.* Overriding imatinib resistance with a novel ABL kinase inhibitor. *Science* **305**, 399–401 (2004).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank R. Lawrence and C. Field for technical assistance with the BCR-ABL quantitative and mutation analysis. N.P.S. and C.L.S. acknowledge support from the Leukemia and Lymphoma Society. The Program for Evolutionary Dynamics at Harvard University is supported by J. Epstein. N.P.S. is the recipient of a Mentored Clinical Pharmacology Research Scholars Program Award.

Author Contributions F.M. is a Junior Fellow of the Harvard Society of Fellows. C.L.S. is an Investigator of the Howard Hughes Medical Institute, and a Doris Duke Distinguished Clinical Scientist.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to F.M. (michor@fas.harvard.edu).

Escherichia coli swim on the right-hand side

Willow R. DiLuzio^{1,2}, Linda Turner³, Michael Mayer¹, Piotr Garstecki¹, Douglas B. Weibel¹, Howard C. Berg^{3,4} & George M. Whitesides¹

The motion of peritrichously flagellated bacteria close to surfaces is relevant to understanding the early stages of biofilm formation and of pathogenic infection^{1–4}. This motion differs from the random-walk trajectories⁵ of cells in free solution. Individual *Escherichia coli* cells swim in clockwise, circular trajectories near planar glass surfaces^{6,7}. On a semi-solid agar substrate, cells differentiate into an elongated, hyperflagellated phenotype and migrate cooperatively over the surface⁸, a phenomenon called swarming. We have developed a technique for observing isolated *E. coli* swarmer cells⁹ moving on an agar substrate and confined in shallow, oxidized poly(dimethylsiloxane) (PDMS) microchannels. Here we show that cells in these microchannels preferentially ‘drive on the right’, swimming preferentially along the right wall of the microchannel (viewed from behind the moving cell, with the agar on the bottom). We propose that when cells are confined between two interfaces—one an agar gel and the second PDMS—they swim closer to the agar surface than to the PDMS surface (and for much longer periods of time), leading to the preferential movement on the right of the microchannel. Thus, the choice of materials guides the motion of cells in microchannels.

Peritrichously flagellated bacteria are propelled by long (about 10 μm), thin, helical filaments distributed randomly over the surface of the cell body. A reversible rotary motor embedded in the cell wall drives each filament at its base^{10,11}. If all motors are rotating anticlockwise, the flagella bundle together and propel the cell forward in a ‘run’. When one or more of the motors switches to clockwise rotation, the corresponding flagella unbundle and reorient the cell in a ‘tumble’¹². During a run, the forward thrust generated by the flagellar bundle is balanced by the viscous drag on the cell body, and the torque produced by the rotating flagellar bundle is balanced by the torque due to the counter-rotation of the cell body¹³. If a cell swims close to a planar surface, these rotations and the resistance from the surface affect the direction of movement. The flagellar bundle rolls to the left near the surface, and the cell body rolls to the right near the surface. These two motions cause the cell to swim in a clockwise, circular trajectory^{4,6,7,14}.

Cells swim in circles at surfaces for seconds to minutes, although one might expect them to drift from the surface quickly because of the effects of rotational brownian motion and bundle fluctuation (wobble) on their trajectories^{5,15}. Hydrodynamic interactions cause the extended interaction of cells with surfaces^{4,14}. Figure 1 shows a schematic representation of *E. coli* cells swimming near two horizontal surfaces. The cells swim in clockwise, circular trajectories at each surface: the trajectories of cells close to the bottom surface seem to follow clockwise paths, and the trajectories of cells close to the top surface seem to follow anticlockwise paths^{6,7,15}. (A clockwise trajectory appears anticlockwise when viewed from the opposite side.) The direction of curvature of the trajectory of the cell therefore indicates whether the cells are swimming closer to the top surface or to the bottom surface.

We have developed a new technique for examining the movement of individual bacteria on nutrient agar by confining the cells in shallow microchannels to constrain their motion to two dimensions. Using soft lithography¹⁶, we fabricated thin (150 μm thick), flexible, gas-permeable¹⁷ films of PDMS embossed with grooves. The surface of the film was rendered hydrophilic by treatment with an air plasma (the advancing contact angle of water on the PDMS film after treatment was 10–20°). We placed the oxidized PDMS (ox-PDMS) film on the agar a few millimetres from the edge of a swarm of *E. coli*⁹. The film sealed conformally to the agar substrate and formed microchannels in which the bottom agar surface formed the floor of the channel, and the ox-PDMS film formed the sidewalls and ceiling (Fig. 2a). An aqueous solution of nutrients from the agar substrate wetted and filled the hydrophilic microchannels. Observation of small, suspended polystyrene beads in the channel showed that no net flow of fluid occurred in the microchannels once they had filled. Individual cells from the swarm migrated into the microchannels. Once they entered the liquid-filled microchannels, they separated from other swarmer cells and swam independently.

In rectangular agar/ox-PDMS microchannels that were only slightly taller than the width of one cell (1.3–1.5 μm tall and 7–10 μm wide), most *E. coli* cells moving in either direction swam preferentially along the channel wall to their right (when viewed from above). The swimming of cells in a clockwise direction (movement to the right) implies that cells are swimming closer to the bottom surface (agar) than to the top surface (ox-PDMS) (Fig. 1). This ordered movement to the right in opposite directions along the two walls resembles that of cars driving along a two-way street. Figure 2b–d

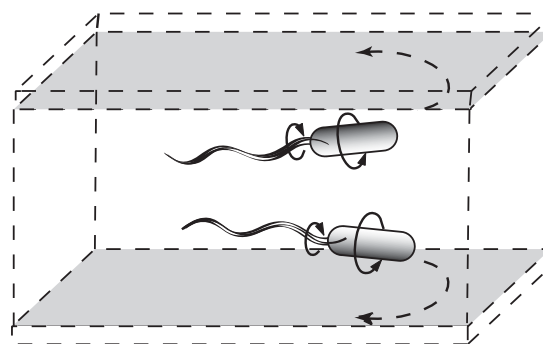


Figure 1 | Cells swim in clockwise, circular trajectories at solid, planar surfaces. When a cell executes a run, all flagella in the bundle rotate anticlockwise (when viewed from behind) and the cell body counter-rotates in a clockwise direction. When viewed from above, the cell trajectories at the bottom surface appear clockwise and the cell trajectories at the top surface appear anticlockwise.

¹Department of Chemistry and Chemical Biology, Harvard University, 12 Oxford Street, Cambridge, Massachusetts 02138, USA. ²Division of Engineering and Applied Sciences, Harvard University, 29 Oxford Street, Cambridge, Massachusetts 02138, USA. ³The Rowland Institute at Harvard, 100 Edwin H. Land Boulevard, Cambridge, Massachusetts 02142, USA. ⁴Department of Molecular and Cellular Biology, Harvard University, 16 Divinity Avenue, Cambridge, Massachusetts 02138, USA.

shows time-lapse images of the traffic-like behaviour of cells in microchannels (see Supplementary Movies).

We wanted to confirm that the motion of cells in microchannels was analogous to the motion of cells confined between two glass surfaces; that is, that cells moving clockwise (or to the right) were at the bottom surface, and that cells moving anticlockwise (or to the left) were at the top surface. We could not easily observe the surface to which the cells were closest in agar/ox-PDMS channels 1.3–1.5 μm tall; we then examined taller channels. Using fluorescence microscopy, we observed HCB437 cells, expressing enhanced green fluorescent protein, in rectangular (5 μm tall and 10 μm wide) ox-PDMS channels sealed to glass. Imaging with a 100 $\times$, 1.4 numerical aperture, oil objective, we could easily focus on either the floor or the ceiling of the microchannel and directly observe the surface to which cells were closest. Cells swimming near the floor of the channel were moving along the channel wall on their right (with respect to the direction of movement of the cells); cells swimming near the ceiling of the channel were moving along the channel wall on their left.

Although most cells swam on the right in the 1.3–1.5 μm tall agar/ox-PDMS microchannels, occasionally some cells travelled in the 'wrong' direction, swimming on the left. To quantify the preference of cells to swim on the right (or to swim closer to the agar surface than the ox-PDMS surface), we fabricated ox-PDMS films embossed with microchannels containing three-way junctions (Fig. 3). Swarmer

cells travelling along the right channel wall entered the right side of the junction; cells travelling along the left channel wall entered the left side of the junction. We counted those cells that passed through the junction and entered each curving side-channel. We defined the percentage of cells swimming on the right as the number of cells that entered the right channel divided by the total number of cells that entered the left or right channels. We did not count the small number of cells that continued straight and did not enter either side channel. Table 1 shows the resulting preference of cells to swim on the right for different bacterial strains and materials comprising the channels.

Smooth-swimming *E. coli* cells, which do not tumble, exhibited a stronger preference to travel on the right side of the channel than did wild-type cells of *E. coli*. Smooth-swimming cells stayed aligned with the channel wall to their right over distances of several millimetres; when wild-type cells tumbled in the rectangular channels, they briefly lost their preference for the right-hand wall. One of the trajectories in Fig. 2e shows a wild-type cell that tumbled and temporarily moved away from the microchannel wall. Even smooth-swimming cells did not swim along the right wall indefinitely. Wobbling or rotational brownian motion eventually caused cells to separate from the wall and then to reassociate (apparently randomly) with either the left or right channel wall. The preference to the right in shallow agar/ox-PDMS channels for all strains examined indicates that tumbling, wobbling and/or rotational brownian motion is suppressed more when cells swim near the agar surface than when cells swim near the PDMS surface.

Cells continued to swim close to the agar surface even when we inverted the experimental system. In an inverted system, nutrient agar formed the channel ceiling and ox-PDMS formed the walls and floor; here cells preferred to move to the left (when viewed from above); that is, they swam closer to the top agar surface. This observation shows that the preferential movement of cells for the right when agar is the floor of the microchannel is not a result of the influence of gravity. We observed preferential movement of the cells only when agar was used as either the floor or ceiling of a composite microchannel; when the floor of the channel was composed of oxidized glass or ox-PDMS, and the channel sidewalls and ceiling

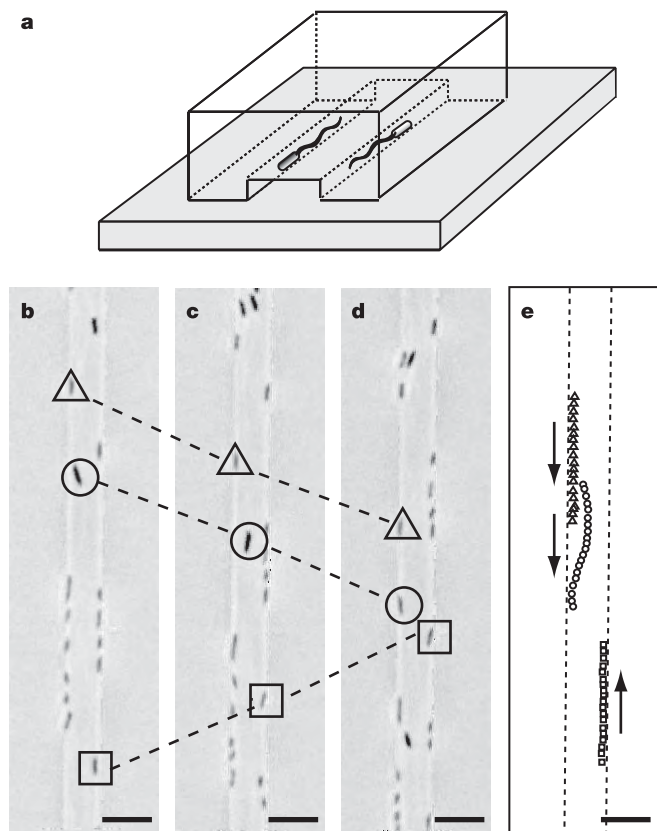


Figure 2 | Images of cells in composite agar/ox-PDMS microchannels. **a**, An oxidized PDMS film with embossed channels seals conformally to an agar substrate and forms liquid-filled channels into which *E. coli* swarmer cells migrate. Within these microchannels (imaged from above), cells that are closer to the floor of the channel swim clockwise or move to the right. **b–d**, Three time-lapse images showing *E. coli* swarmer (AW405) cells²⁶ moving on the right in confining, rectangular microchannels (1.4 μm tall and 7 μm wide) in which nutrient agar formed the floor of the channel and an ox-PDMS film formed the walls and ceiling of the channel. **b**, $t = 0$ s; **c**, $t = 0.33$ s; **d**, $t = 0.67$ s. **e**, The trajectories of three selected cells between images **b** and **d** are shown. Scale bars, 10 μm .

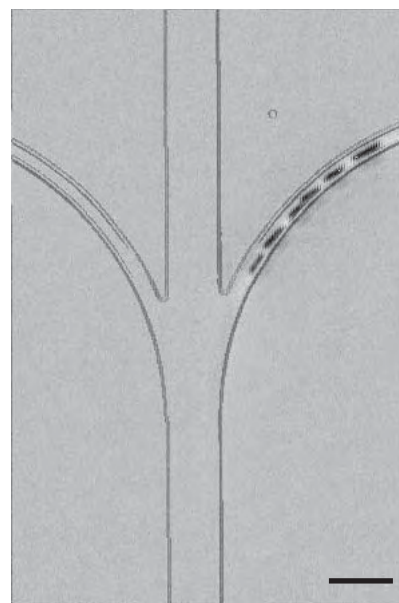


Figure 3 | Quantification of cells displaying a preference to travel to the right by using a microchannel junction. Cells travelling along the right channel wall (closer to the floor of the microchannel) enter the right sidearm; cells travelling along the left channel wall (closer to the ceiling of the microchannel) enter the left sidearm. Six smooth-swimming *E. coli* (HCB437) cells²⁷ are shown entering the right sidearm. Scale bar, 20 μm .

Table 1 | Determination of preference of cells for swimming on the right

<i>E. coli</i> strain	Floor material	Ceiling material	Cells swimming on the right (%)
AW405	Nutrient agar	Ox-PDMS	75 ± 7
RP437	Nutrient agar	Ox-PDMS	77 ± 2
HCB437	Nutrient agar	Ox-PDMS	88 ± 7
HCB437	Ox-PDMS	Nutrient agar	16 ± 13
HCB437	Ox-PDMS	Ox-PDMS	55 ± 12
HCB437	Ox-glass	Ox-PDMS	53 ± 2

The microchannels used for analysis were 1.3–1.5 μm tall and 7–10 μm wide. Strains AW405 (ref. 26) and RP437 (ref. 28) are wild-type for chemotaxis, and strain HCB437 (ref. 27) is a smooth-swimming strain that is deleted for most chemotaxis genes. The percentage of cells swimming on the right was determined as 100 times the number of cells that entered the right divided by the sum of the cells that entered the right and the cells that entered the left. For each percentage given, at least 1,000 cells were counted in at least seven separate channels. Errors are s.d. for the individual channel percentages. Nutrient agar contained 3 g l⁻¹ beef extract, 10 g l⁻¹ peptone, 5 g l⁻¹ sodium chloride, 4.5 g l⁻¹ Eiken agar and 0.5% glucose.

were made of ox-PDMS, and channels were filled with liquid growth medium, cells showed no preference for one side of the microchannel.

To examine whether short-range interactions between the bacteria and the surfaces of the microchannels contributed to the mechanism underlying the preference for the right, we added surfactants to the solution used to cast the agar that formed the floor of the microchannel; we also changed its ionic strength. For these studies we used non-nutrient motility agar (10 mM potassium phosphate pH 7.0, 0.5% Eiken agar). The addition of a surface-active agent such as bovine serum albumin or surfactin (a lipopeptide biosurfactant produced by *B. subtilis*¹⁸) to the motility agar was necessary to inhibit the adhesion of cells to the ox-PDMS surfaces, but we continued to observe a preferential movement of cells to the right (more than 80% preference for all the surfactants examined; see Supplementary Information). Each surfactant would be expected to alter the van der Waals and steric interaction of cells with the channel walls in a slightly different way. We also varied the ionic strength of the motility agar, and continued to observe a preference of more than 80% for the right (see Supplementary Information). Taken together, these results indicate that short-range molecular interactions (van der Waals, ionic, hydrogen bond, hydrophobic or steric) are not relevant for the preferential movement of cells.

As the channel height increased, the preference of cells for movement along the right wall slowly diminished (Fig. 4). The simplest

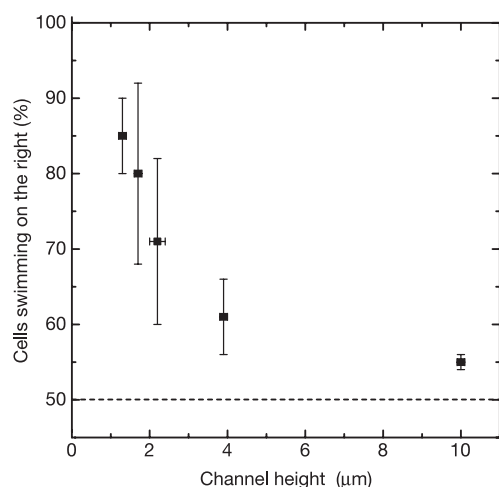


Figure 4 | Preference of cells to move on the right in microchannels as a function of the height of the channel. Channels had a floor composed of nutrient agar, and a ceiling and sidewalls composed of ox-PDMS. For each datum, at least 10³ cells were counted in seven or more separate channels. Error bars show s.d. for the different channels measured.

rationale for the slowly decaying preference to the right with channel height is that cells must experience different hydrodynamic environments when they swim near agar surfaces and ox-PDMS surfaces, because hydrodynamic interactions fall off slowly with distance. When the channel height was 10 μm , about equal numbers of cells swam to the right (closer to the microchannel floor) and to the left (closer to the microchannel ceiling); that is, there was no preference for side. These results are consistent with other studies indicating that cells are affected by a surface (here, the gel) only when they swim within 10 μm of the surface^{6,19,20}.

Agar and agarose are porous gels with a wide distribution of pore sizes²¹. We propose that cells experience less resistance to their movement when they swim close to the porous agar surface than when they swim close to the non-porous ox-PDMS surface. It has been shown that a translating and/or rotating sphere experiences less hydrodynamic drag near a porous boundary than near a solid boundary²². The speed at which a cell swims provides a direct indicator of the hydrodynamic drag it experiences. We measured the swimming speed of 50 individual cells (strain HCB437) in channels with porous agar floors and in channels with solid ox-PDMS floors (Table 2). Cells moving on the right in channels with agar floors swam at an average speed of $31 \pm 3 \mu\text{m s}^{-1}$ and cells in channels with ox-PDMS floors swam at $27 \pm 4 \mu\text{m s}^{-1}$; these mean velocities were statistically different within a 95% confidence limit. These results further suggest that the hydrodynamics of swimming is different in composite agar/ox-PDMS microchannels and in microchannels in which all of the walls are ox-PDMS.

To determine whether there were phenotypic differences between cells that swam along the left or the right wall in shallow agar/ox-PDMS channels, we compared the lengths and swimming speeds of cells at each wall. In an agar/ox-PDMS channel that was 1.4 μm tall and 10 μm wide there was no significant difference in the average length of cells travelling along the right or left sidewall (Table 2). However, cells moving along the right wall swam faster (by 15%) than cells moving along the left, which is consistent with the hypothesis that cells experience less resistance when they swim closer to the agar surface than to the PDMS surface. In an ox-PDMS/ox-PDMS channel that was 1.4 μm tall and 10 μm tall, there was no significant difference in the average length or speed of the cells travelling on the left or right wall (Table 2).

The finding that cells move preferentially along the surface of polysaccharide hydrogel for a longer period than along a solid surface might help to rationalize behaviours that are important in environmental and medical microbiology. Various bacteria often produce polysaccharides to enhance flagella-dependent swarming motility, pili-dependent twitching motility, and gliding motility on surfaces². Uropathogenic *Proteus mirabilis* cells, which are often of the swarming phenotype, migrate up the urinary tract, which is lined with polysaccharide-coated uroepithelial cells²³. *Vibrio fischeri* symbiotically colonize the light organ of the squid *Euprymna scolopes* by migrating through narrow ducts (15 μm wide) that are coated with mucus produced by the squid²⁴.

Table 2 | Length and swimming speeds of cells travelling on right versus left

Floor material	Ceiling material	Side of channel	Cell length (μm)	Swimming speed ($\mu\text{m s}^{-1}$)
Nutrient agar	Ox-PDMS	Right	3.3 ± 0.8	31 ± 3
		Left	3.4 ± 0.9	27 ± 4
Ox-PDMS	Ox-PDMS	Right	2.7 ± 0.7	26 ± 5
		Left	2.6 ± 0.4	26 ± 5

Lengths and swimming speeds were measured for smooth-swimming HCB437 cells²⁷ in microchannels that were 1.3–1.5 μm tall and 10 μm wide; $n = 50$ cells for each determination. Errors are s.d. for 50 measurements. Histograms showing the distribution of the lengths of cells and swimming speeds are given in Supplementary Information. Nutrient agar contained 3 g l⁻¹ beef extract, 10 g l⁻¹ peptone, 5 g l⁻¹ sodium chloride, 4.5 g l⁻¹ Eiken agar and 0.5% glucose.

The composite agar/ox-PDMS channels developed here can be extended to study the growth or movement of a range of motile, free-swimming microorganisms (or perhaps higher organisms) in a confined but nutritive environment. Observing a preference for swimming along the right or left wall provides an experimentally convenient method for examining the hydrodynamics of movement close to proximal surfaces, and to determine the chemical and physical dependence of these interactions. The ability to design channels by using materials that affect the hydrodynamics of swimming differently offers a strategy with which to direct the motion of motile cells in microchannels. This strategy would not require external pumping, valving or chemical gradients to direct the movement of cells. We believe that directed motion is a first step towards the development of self-contained microdevices that use moving bacterial cells²⁵ for cell-based bioassays and biosensors.

Received 4 February; accepted 21 April 2005.

- Pratt, L. A. & Kolter, R. Genetic analysis of *Escherichia coli* biofilm formation: Roles of flagella, motility, chemotaxis and type I pili. *Mol. Microbiol.* **30**, 285–293 (1998).
- Harshey, R. M. Bacterial motility on a surface: Many ways to a common goal. *Annu. Rev. Microbiol.* **57**, 249–273 (2003).
- Ottemann, K. M. & Miller, J. F. Roles for motility in bacterial–host interactions. *Mol. Microbiol.* **24**, 1109–1117 (1997).
- Vigeant, M. A. S., Ford, R. M., Wagner, M. & Tamm, L. K. Reversible and irreversible adhesion of motile *Escherichia coli* cells analyzed by total internal reflection aqueous fluorescence microscopy. *Appl. Environ. Microbiol.* **68**, 2794–2801 (2002).
- Berg, H. C. & Brown, D. A. Chemotaxis in *Escherichia coli* analyzed by three-dimensional tracking. *Nature* **239**, 500–504 (1972).
- Berg, H. C. & Turner, L. Chemotaxis of bacteria in glass-capillary arrays. *Biophys. J.* **58**, 919–930 (1990).
- Frymier, P. D., Ford, R. M., Berg, H. C. & Cummings, P. T. Three-dimensional tracking of motile bacteria near a solid planar surface. *Proc. Natl Acad. Sci. USA* **92**, 6195–6199 (1995).
- Henrichsen, J. Bacterial surface translocation: A survey and classification. *Bacteriol. Rev.* **36**, 478–503 (1972).
- Harshey, R. M. & Matsuyama, T. Dimorphic transition in *Escherichia coli* and *Salmonella typhimurium*—surface-induced differentiation into hyperflagellate swarmer cells. *Proc. Natl Acad. Sci. USA* **91**, 8631–8635 (1994).
- Berg, H. C. & Anderson, R. A. Bacteria swim by rotating their flagellar filaments. *Nature* **245**, 380–382 (1973).
- Berg, H. C. The rotary motor of bacterial flagella. *Annu. Rev. Biochem.* **72**, 19–54 (2003).
- Turner, L., Ryu, W. S. & Berg, H. C. Real-time imaging of fluorescent flagellar filaments. *J. Bacteriol.* **182**, 2793–2801 (2000).
- Taylor, G. The action of waving cylindrical tails in propelling microscopic organisms. *Proc. R. Soc. Lond. A* **211**, 225–239 (1952).
- Ramias, M., Tullock, D. L. & Phan-Thien, N. The role of hydrodynamic interaction in the locomotion of microorganisms. *Biophys. J.* **65**, 755–778 (1993).
- Vigeant, M. A. S. & Ford, R. M. Interactions between motile *Escherichia coli* and glass in media with various ionic strengths, as observed with a three-dimensional-tracking microscope. *Appl. Environ. Microbiol.* **63**, 3474–3479 (1997).
- Xia, Y. & Whitesides, G. M. Soft lithography. *Angew. Chem. Int. Edn Engl.* **37**, 550–575 (1998).
- Brandrup, J., Immergut, E. H. & Grulke, E. A. (eds) *Polymer Handbook*, 4th edn (Wiley, New York, 1999).
- Toguchi, A., Siano, M., Burkart, M. & Harshey, R. M. Genetics of swarming motility in *Salmonella enterica* Serovar Typhimurium: Critical role for lipopolysaccharide. *J. Bacteriol.* **182**, 6308–6321 (2000).
- Frymier, P. D. & Ford, R. M. Analysis of bacterial swimming speed approaching a solid–liquid interface. *AIChE J.* **43**, 1341–1347 (1997).
- Biondi, S. A., Quinn, J. A. & Goldfine, H. Random motility of swimming bacteria in restricted geometries. *AIChE J.* **44**, 1923–1929 (1998).
- Pernodet, N., Maaloum, M. & Tinland, B. Pore size of agarose gels by atomic force microscopy. *Electrophoresis* **18**, 55–58 (1997).
- Damiano, E. R., Long, D. S., El-Khatib, F. H. & Stace, T. M. On the motion of a sphere in a Stokes flow parallel to a Brinkman half-space. *J. Fluid Mech.* **500**, 75–101 (2004).
- Allison, C., Lai, H. C. & Hughes, C. Coordinate expression of virulence genes during swarm-cell differentiation and population migration of *Proteus mirabilis*. *Mol. Microbiol.* **6**, 1583–1591 (1992).
- Visick, K. L. & McFall-Ngai, M. J. An exclusive contract: Specificity in the *Vibrio fischeri* *Euprymna scolopes* partnership. *J. Bacteriol.* **182**, 1779–1787 (2000).
- Darnton, N., Turner, L., Breuer, K. & Berg, H. C. Moving fluid with bacterial carpets. *Biophys. J.* **86**, 1863–1870 (2004).
- Armstrong, J. B., Adler, J. & Dahl, M. M. Nonchemotactic mutants of *Escherichia coli*. *J. Bacteriol.* **93**, 390–398 (1967).
- Wolfe, A. J., Conley, M. P., Kramer, T. J. & Berg, H. C. Reconstitution of signalling in bacterial chemotaxis. *J. Bacteriol.* **169**, 1878–1885 (1987).
- Parkinson, J. S. Complementation analysis and deletion mapping of *Escherichia coli* mutants defective in chemotaxis. *J. Bacteriol.* **135**, 45–53 (1978).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank W. S. Ryu, D. Ryan, M. P. Brenner and H. A. Stone for discussions, and S. Rojevskaya for technical assistance. This research was supported by the NIH and DOE. W.R.D. acknowledges an NSF-IGERT Biomechanics Training Grant. M.M. acknowledges a postdoctoral fellowship from the Swiss National Science Foundation. P.G. thanks the Foundation for Polish Science for a postdoctoral fellowship. D.B.W. thanks the NIH for a postdoctoral fellowship.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to G.M.W. (gwhitesides@gmwgroup.harvard.edu).

RNA-interference-directed chromatin modification coupled to RNA polymerase II transcription

Vera Schramke¹, Daniel M. Sheedy¹, Ahmet M. Denli², Carolina Bonila³, Karl Ekwall³, Gregory J. Hannon² & Robin C. Allshire¹

RNA interference (RNAi) acts on long double-stranded RNAs (dsRNAs) in a variety of eukaryotes to generate small interfering RNAs that target homologous messenger RNA, resulting in their destruction. This process is widely used to 'knock-down' the expression of genes of interest to explore phenotypes^{1–3}. In plants^{3–5}, fission yeast^{6–8}, ciliates^{9,10}, flies¹¹ and mammalian cells^{12,13}, short interfering RNAs (siRNAs) also induce DNA or chromatin modifications at the homologous genomic locus, which can result in transcriptional silencing or sequence elimination¹⁴. siRNAs may direct DNA or chromatin modification by siRNA–DNA interactions at the homologous locus^{4,5}. Alternatively, they may act by interactions between siRNA and nascent transcript^{15,16}. Here we show that in fission yeast (*Schizosaccharomyces pombe*), chromatin modifications are only directed by RNAi if the homologous DNA sequences are transcribed. Furthermore, transcription by exogenous T7 polymerase is not sufficient. Ago1, a component of the RNAi effector RISC/RITS complex, associates with target transcripts and RNA polymerase II. Truncation of the regulatory carboxy-terminal domain (CTD) of RNA pol II disrupts transcriptional silencing, indicating that, like other RNA processing events^{17–19}, RNAi-directed chromatin modification is coupled to transcription.

In plant and mammalian cells siRNAs homologous to the open reading frame of a gene results in post-transcriptional silencing, degrading transcripts by means of RNAi. However, siRNAs homologous to a gene's promoter can induce transcriptional silencing, resulting in the modification of DNA and/or chromatin^{3–5}. siRNAs may hybridize to DNA and thereby recruit DNA/chromatin modifying activities that effect silencing^{4,5,14}. Alternatively, the RNAi machinery may target nascent transcripts and cause chromatin modification on templates homologous to loaded siRNAs^{15,16}. At fission yeast centromeres and the silent mating type locus, non-coding RNAs are generated by the transcription of both strands of related repeats^{6,20}. These form dsRNAs, which are cleaved by Dicer (Dcr1) into siRNAs and then loaded into the Ago1 (Argonaute)-containing RITS (for RNA-induced initiation of transcriptional silencing) complex, which mediates RNAi^{20,21}. Nascent transcripts may direct the RNAi machinery to the homologous locus, induce dimethylation of the surrounding chromatin on lysine 9 of histone H3 (H3K9me2) through Clr4, recruiting Swi6 (HP1) and thereby silencing transcription^{15,16}. Components of RITS and RNA-dependent RNA polymerase (Rdp1) are known to associate with, and act in *cis* on, this silent chromatin^{6,21,22}. However, Chp1 (a RITS component) and Swi6 bind H3K9me2 (ref. 23), and Rdp1 (and the RDRC (RNA-directed RNA polymerase complex)) interacts with RITS and requires Swi6 for chromatin association^{24,25}. Because of this and the inherent self-enforcing nature of the process, it is difficult to

determine whether nascent transcripts are required to mediate RNAi-directed chromatin modifications, and what additional interactions are involved.

Expression of a synthetic hairpin RNA homologous to a 280-base-pair (bp) region located within the *ura4*⁺ gene (*sh-ura4-280*) induces Dicer-dependent transcriptional silencing of *ura4*⁺ along with H3K9me2 of *ura4* chromatin and recruitment of Swi6 (ref. 7). To determine whether this hairpin-induced chromatin modification requires a homologous transcript, a strain containing a modified *ura4*⁺ gene with an efficient transcription terminator module immediately upstream of the 280-bp *ura4* target region was used²⁶. In this strain (Ter⁺ *ura4*) the transcriptional terminator is inserted within an artificial intron, so that more than 99% of transcripts are terminated before the 280-bp region homologous to the *sh-ura4-280* trigger. A second strain (Ter-M5 *ura4*) contains a *cis*-acting mutation within the terminator, allowing 75% of transcripts to traverse the downstream 280-bp region of the gene²⁶ (Fig. 1a). Both strains also contain *ura4-DS/E* at the *ura4*⁺ locus, which is fully transcribed but lacks the 280-bp region homologous to the *sh-ura4-280* trigger, thus providing a convenient internal control⁷.

The construct expressing *sh-ura4-280* was introduced into both strains, and transcription of *ura4*⁺, H3K9me2 modification of *ura4* chromatin and recruitment of Swi6 were assessed in wild-type strains in the presence and absence of the *sh-ura4-280* hairpin. In cells containing the Ter⁺ *ura4* gene, truncated (*ura-T*), but not full-length (*ura-FL*), transcript is detected in the presence or absence of *sh-ura4-280* (Fig. 1b). In Ter-M5 cells, full-length *ura4* transcript is lost in the presence of *sh-ura4-280* but expression of *ura4-DS/E* remains unaffected (Fig. 1b). Thus, expression of a hairpin target homologous to downstream DNA sequences does not affect Ter⁺ *ura4*, whereas Ter-M5 *ura4* transcripts are repressed by *sh-ura4-280* expression.

Chromatin immunoprecipitation (ChIP) was used to assess the levels of H3K9me2 modification on the Ter⁺ and Ter-M5 *ura4* genes relative to *ura4-DS/E*. H3K9me2 was detected only on Ter-M5 *ura4*, and only in strains expressing *sh-ura4-280* (Fig. 1c). *ura4* siRNAs were detected only in strains containing the *sh-ura4-280* construct (Fig. 1d). Thus, RNAi can induce chromatin modifications at a homologous locus only if transcripts traverse a region identical in sequence to the hairpin trigger and the resultant siRNAs. It is possible that the passage of RNA polymerase II (pol II) during transcription itself, by opening chromatin, provides access for siRNAs to underlying DNA sequences, thus allowing siRNA–DNA interactions^{4,5,14}. Alternatively, Ago1-bearing siRNAs may bind homologous nascent transcripts and in so doing recruit chromatin-modifying activities through the Ago1-containing RITS and/or Rdp1-containing RDRC complex^{15,16}. If opening the two DNA strands is sufficient, then an

¹Wellcome Trust Centre for Cell Biology, Institute of Cell and Molecular Biology, King's Buildings, University of Edinburgh, Edinburgh EH9 3JR, UK. ²Cold Spring Harbor Laboratory, Watson School of Biological Sciences, 1 Bungtown Road, Cold Spring Harbor, New York 11724, USA. ³Karolinska Institutet, Department of Biosciences/Department of Natural Sciences, University College Södertörn, Alfred Nobel's Allé 7, S-141 89 Huddinge, Sweden.

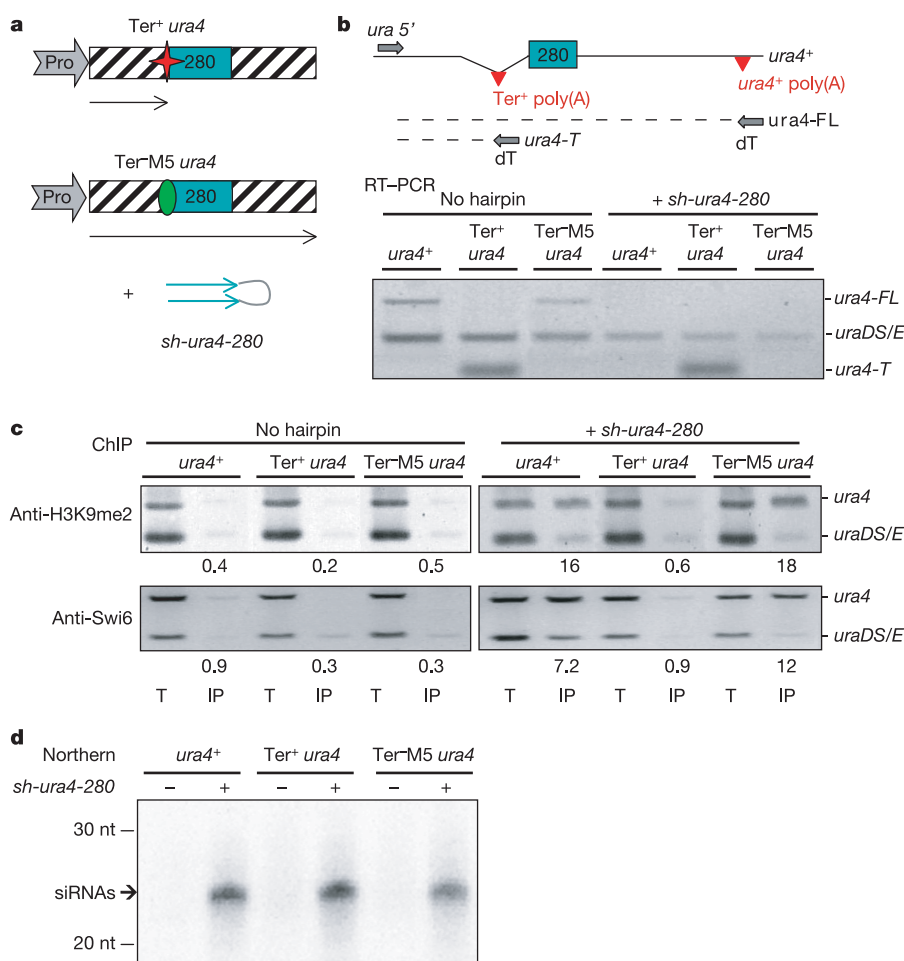


Figure 1 | Transcription of siRNA target is required to effect silent chromatin assembly. **a**, Strains containing the *Ter⁺-ura4* gene with an efficient *Ter⁺* or defective *TerM5* terminator upstream of the 280-bp *ura4* target region. **b**, RT-PCR analysis of *ura4⁺* transcripts on oligo(dT)-primed cDNA from RNA samples of the indicated strains; expression of *ura4⁺*-terminated (*ura-T*) and full-length (*ura4-FL*) transcripts relative to the *ura4-DS/E* minigene in the same strains with or without *sh-ura4-280* hairpin. **c**, ChIP analyses with anti-H3K9me2 and Swi6 antibodies over the *ura4* gene in the indicated strains. T, total extract; IP, immunoprecipitate. **d**, Detection of *sh-ura4-280*-generated siRNAs by northern blotting. nt, nucleotides.

exogenous RNA polymerase might allow siRNAs access to homologous chromatin. To test this, the *ura4* transcription unit was placed downstream of the bacteriophage T7 promoter (Fig. 2). T7 polymerase was constitutively expressed from the *adh1* promoter in the presence or absence of *sh-ura4-280* (Fig. 2a, and Supplementary Fig. S1a). *T7:ura4* transcripts are detected only in cells expressing T7 polymerase. The expression of *sh-ura4-280* did not reduce the level of *T7:ura4* transcripts significantly (Fig. 2b), although RNAi is active because *ura4* siRNAs are readily detected (Fig. 2d). In addition, no H3K9me2 or Swi6 could be detected on *T7:ura4* chromatin (Fig. 2c, and Supplementary Fig. S1b) in cells expressing *sh-ura4-280* homologous siRNAs (Fig. 2d), although histone H3 is present on the *T7:ura4* gene (Supplementary Fig. S1c). Lack of RNAi-directed chromatin modification of the *T7:ura4* template may reflect the absence of features normally associated with endogenous RNA pol II transcription. T7 and RNA pol II transcription and the resulting transcripts differ in many respects; regardless of this, transcription of target chromatin alone is not sufficient to mediate RNAi-directed chromatin modifications on homologous chromatin. This indicates that transcripts generated by, or associated with, a specific RNA polymerase might be required.

RNA pol II is responsible for the generation of fission yeast centromere repeat transcripts that are processed by RNAi into homologous siRNAs. A mutation (*rpb7-1*) in Rpb7, a small pol II subunit, leads to a loss of these transcripts and siRNAs (K.E., unpublished observations). The CTD of the large subunit of pol II (Rpb1) contains multiple conserved YSPTSPS heptad repeats, the phosphorylation state of which regulates the binding of various mRNA processing factors, thus coupling mRNA processing to transcription^{17–19}. In *Saccharomyces cerevisiae* the deletion of up to

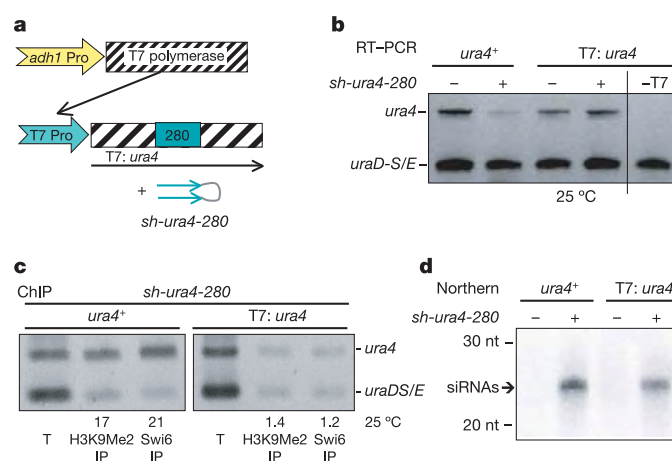


Figure 2 | Transcription by T7 polymerase is not sufficient for RNAi-directed chromatin modification. **a**, The *ura4⁺* promoter was replaced with the bacteriophage T7 promoter (*T7:ura4*). T7 polymerase was constitutively expressed from the *adh1* promoter in the presence or absence of *sh-ura4-280*. Cells were grown at 25 °C. **b**, RT-PCR analysis of *ura4⁺* transcripts was performed on oligo(dT)-primed cDNA from RNA samples from the indicated strains expressing (+) or not (-) *sh-ura4-280*-generated siRNAs. **c**, ChIP analyses with anti-H3K9me2 and anti-Swi6 antibodies over the *ura4* gene in the indicated strains. T, total extract; IP, immunoprecipitate. **d**, Detection of *sh-ura4-280* generated siRNAs by northern blotting.

16 of the 26 CTD heptad repeats from pol II results in compromised RNA polymerase functions²⁷. If pol II has a specific function in mediating RNAi-mediated chromatin modification, then cells bearing a defective pol II might display aberrant silencing of marker genes at centromeres. To examine this, a strain was constructed with 17 of the 28 CTD heptad repeats deleted and simultaneously epitope-tagged (*rpb1-11*, see Methods; Fig. 3a). This strain was slow-growing but viable at all temperatures tested and was clearly defective in its ability to silence centromeric *ura4*⁺ and *ade6*⁺ markers as revealed by increased growth on plates lacking uracil (–ura) and the appearance of white *ade*⁺ colonies, respectively (Fig. 3b, left). Consistent with this was the detection of increased levels of *cen1:ura4*⁺ transcripts (Fig. 3b, right) and decreased levels of H3K9me2 associated with centromere repeats (Fig. 3c). However, centromeric transcripts do not accumulate appreciably in *rpb1-11* compared with *dcr1Δ*, and centromeric siRNAs are readily detected as in the wild type (Fig. 3d).

This indicates that although RNAi remains active it is unable to induce chromatin modifications efficiently on homologous sequences. The phenotype of *rpb1-11* is clearly distinct from that of *rpb7-1*, which is defective in centromeric transcription and siRNA production, causing a failure of transcriptional silencing (K.E., unpublished observations). Microarray expression profiling indicated that none of the known genes involved in RNAi-directed chromatin silencing are significantly affected in *rpb1-11* cells in comparison with the wild type, and few genes were affected to any great extent (Supplementary Table 1). Thus, the CTD truncation does not seem to cause a substantial general defect in transcription. This indicates that the CTD of pol II might act downstream of RNAi to stabilize interactions between RNAi components, the nascent transcript and possibly the pol II holoenzyme to induce chromatin modifications. RNAi components might require intact pol II to fully engage a chromatin-associated nascent transcript, or intact pol II

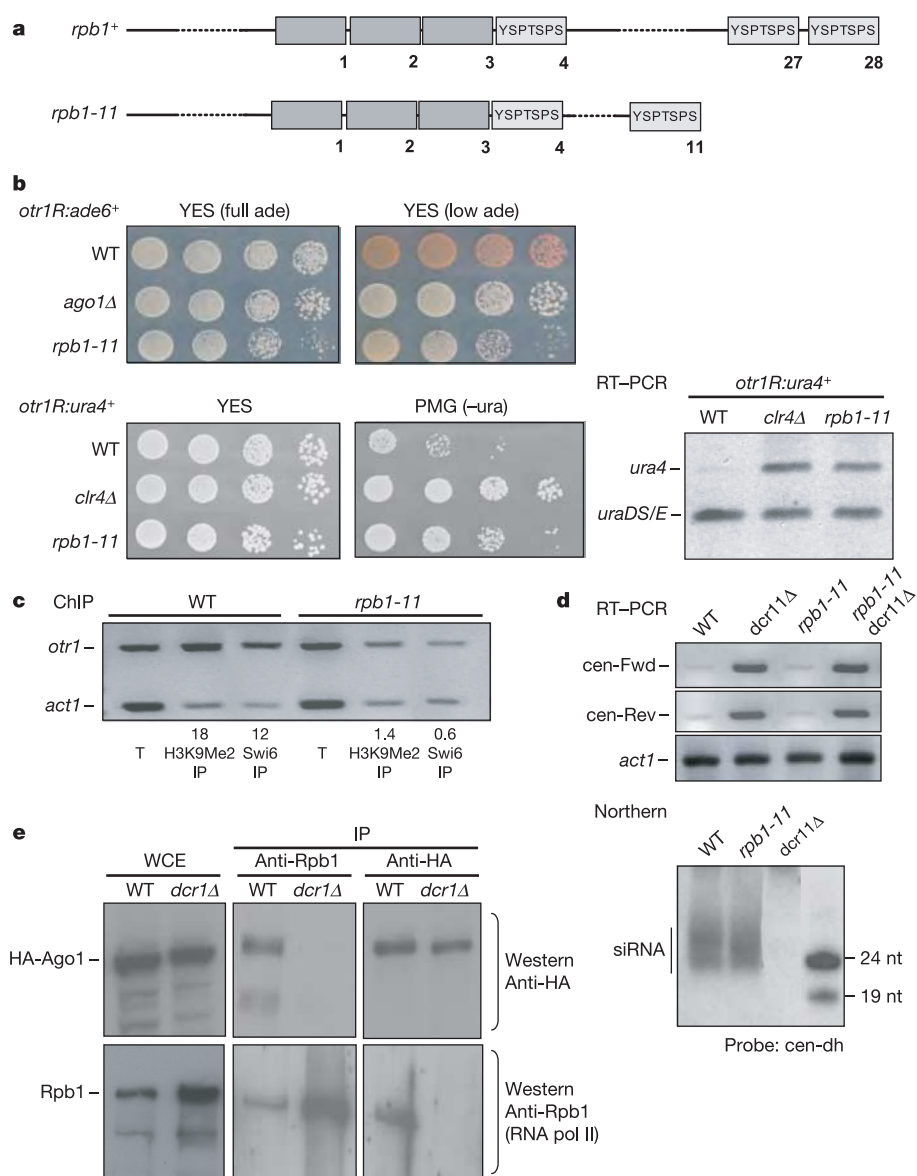


Figure 3 | RNA pol II CTD truncation affects centromeric silent chromatin. **a**, *rpb1-11* retains 11 of 28 heptad repeats. **b**, Left: growth assay of indicated strains with *ade6*⁺ or *ura4*⁺ genes in *cen1* outer repeat on non-selective (YES), limiting adenine (YES (low ade)) or without uracil (PMG (–ura)). Right: RT-PCR analysis of *cen1:ura4*⁺ transcripts performed on RNA from the indicated strains. WT, wild-type. **c**, ChIP

analyses of H3K9me2 and Swi6 over the *cen1* outer repeat region (*otr1*) relative to the actin gene (*act1*). T, total extract; IP, immunoprecipitate. **d**, Detection of centromeric repeat transcripts and siRNAs by RT-PCR (top) and northern blotting (bottom), respectively. **e**, Immunoprecipitation and detection of HA-Ago1 and Rpb1 by western blot analysis.

might be specifically required to synthesize a transcript in a form that can effectively associate with RNAi components. Argonaute (PAZ/PIWI domain) proteins enter RISC (or RITS) complexes and use loaded siRNAs to guide RISC/RITS to target RNAs. Immunoprecipitates of HA-Ago1 were found by western blot analyses to contain pol II (Rpb1); reciprocal to this, HA-Ago1 was detected in immunoprecipitates of pol II. This interaction also required siRNA-loaded RITS because Ago1 and pol II do not immunoprecipitate together from cells lacking Dicer (Fig. 3e).

To determine whether Ago1 associates with chromatin targeted for silencing by RNAi, ChIPs were performed with anti-Ago1 antibodies or HA-Ago1. Ago1 associated with centromeric outer repeats in wild-type, but not *dcrl1Δ*, cells (Fig. 4a, top). Ago1 also showed *sh-ura4-280* siRNA and transcription-dependent association with the *ura4* gene (Fig. 4a, bottom). Ago1, but not Rad21, associated with centromeric *otr* transcripts, but not with control transcripts (*act1*), in wild-type cells, and not in cells lacking siRNAs (*dcrl1Δ*) (Fig. 4b). The association of Ago1 with centromeric chromatin was sensitive to RNase (Fig. 4c). In addition, this association was reduced in strains carrying a truncated pol II CTD (*rpb1-11*) (Fig. 4d). Consistent with previous reports²⁰, immunolocalization shows that HA-Ago1 is concentrated at centromeres in the nucleus, as shown by localization

with centromere-specific CENP-A^{Cnp1} (Fig. 4e, and Supplementary Fig. S2).

Taken together, these data show that, in fission yeast, RNAi requires the transcription of a homologous target to direct chromatin modifications. Opening DNA by T7 pol transcription does not allow modification of the target chromatin to occur. T7 pol might deal with impeding nucleosomes differently, or T7 pol transcripts might not be packaged or processed in the same manner as RNA pol II transcripts, rendering them immune to RNAi. The fact that truncation of the pol II CTD affects RNAi-directed chromatin modifications and association of Ago1 with centromeric repeats, without noticeably affecting centromere repeat transcription or siRNA generation, indicates that pol II transcription might facilitate the conversion of RNAi signals into chromatin modification. Many different factors associate with pol II through its CTD during distinct stages of transcription^{17–19}; the pol II complex might provide a scaffold that promotes interactions between Ago1/RITS-borne siRNA and target pol II transcripts, leading to the efficient modification of occupied chromatin (see model in Supplementary Fig. S3). Indeed, it is known that RNA processing and export seem to be orchestrated with respect to ongoing transcription^{17–19}. Our data indicate that RNAi-directed chromatin modification is another example of an RNA processing event that occurs co-transcriptionally, and offer an explanation for the apparent paradox that RNA pol II is not only required for transcriptional activity but is pivotal in transcriptional silencing. Moreover, plants have even evolved a distinct RNA polymerase (pol IV) required for RNAi-dependent chromatin silencing of certain repeat sequences^{28,29}.

METHODS

Standard techniques. Standard procedures were used for fission yeast growth, genetics and manipulations.

Yeast strains. *S. pombe* strains used are listed in Supplementary Table 2. Deletion of the 17 heptad repeats from Rpb1 and 3 × HA tagging of Ago1 was achieved by standard methods with the indicated primers (Supplementary Table 3). The KAN marker of the 3 × HA Ago1 tag was subsequently swapped to the nourseothricin (NAT) marker with indicated primers (Supplementary Table 3).

Reverse transcriptase polymerase chain reaction (RT-PCR). Total RNA was prepared from strains grown in YES medium at 25 or 32 °C and RT-PCR was performed as described previously⁷. The *ura4* and *ura4-DS/E* PCR products were separated on 1.5% agarose gels and poststained with ethidium bromide. Quantification of bands was performed with the Eastman Kodak EDAS 290 system and 1D Image Analysis software. Analysis was performed two to four times at each temperature, and average values from these experiments are presented.

Western blotting and immunoprecipitations. Antibodies for western blotting were diluted in PBS containing Tween as follows: anti-HA and anti-Rpb1 (monoclonal ARNA-3 from Research Diagnostics detecting residues 797–811 of Rpb1), diluted 1:300; anti-unmodified RNA pol II (8WG16), 1:1,000 (in 5% milk). Blots were developed with enhanced chemiluminescence reagents (Amersham Biosciences). Immunoprecipitations were performed as described previously³⁰.

Chromatin immunoprecipitation. ChIP was performed as described previously⁷ except for the following modifications: cells were converted to spheroplasts by incubation for 25 min at 10⁸ cells ml⁻¹ in PEMS (100 mM PIPES pH 7, 1 mM EDTA, 1 mM MgCl₂, 1.2 M sorbitol) containing 0.4 mg ml⁻¹ zymolyase-100T at 36 °C. Cells were washed twice in PEMS; cell pellets were frozen at -20 °C (ref. 30). Thereafter the standard ChIP procedure was followed. Antibodies against H3K9me2, Swi6 and HA were used as described previously⁷. Quantification of bands was performed with the Eastman Kodak EDAS 290 system and 1D Image Analysis software.

ChIP with an RNase step. ChIPs were performed with a monoclonal antibody against HA (12CA5) and using methods previously described, with the following exceptions. The crosslinking time was reduced from 15 min to 5 min when RNase treatment was performed. When an RNase treatment step was included, cross-linked chromatin from the same experiment was treated with either 7.5 U of RNase A and 300 U of RNase T1 (RNase A/T1 cocktail; Ambion) or an equivalent volume of RNase storage buffer (10 mM HEPES pH 7.2, 20 mM NaCl, 0.1% Triton X-100, 1 mM EDTA, 50% v/v glycerol). After incubation at 25 °C for 30 min, immunoprecipitations were performed as above.

RNA immunoprecipitation. RNA immunoprecipitation was performed as described²⁴.

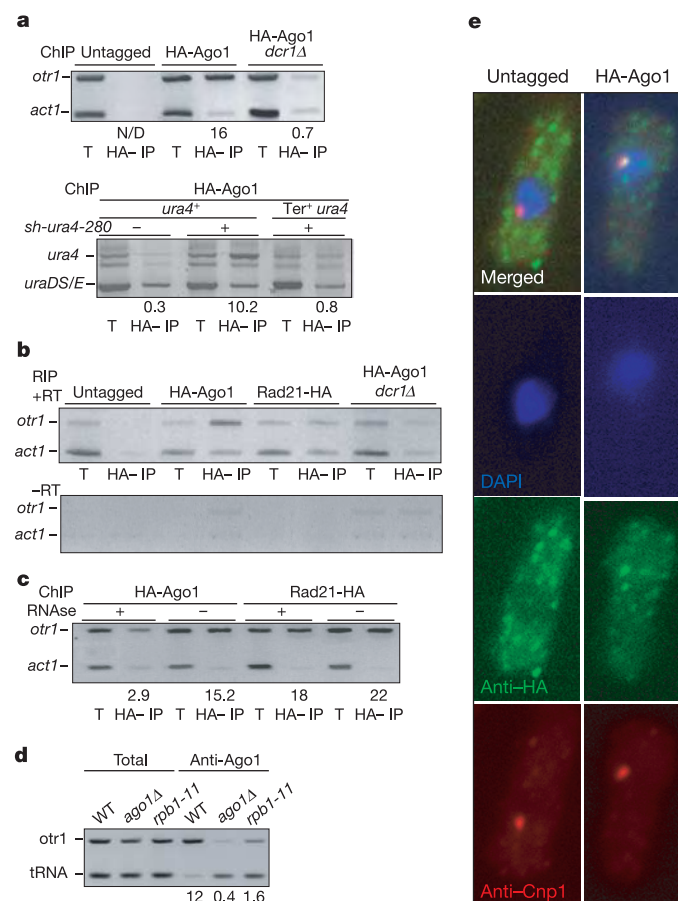


Figure 4 | Association of Ago1 with chromatin is RNase sensitive and dependent on transcription and pol II. **a**, ChIP detecting HA-Ago1 or Rad21-HA on *cen1* outer repeat (top) or *ura4*⁺ with (+) or without (-) *sh-ura4-280* (bottom). **b**, RT-PCR of anti-HA immunoprecipitated RNA to detect HA-Ago1-associated and Rad21-HA-associated centromeric (*otr1*) versus actin (*act1*) transcripts. **c**, ChIP analyses of HA-Ago1 or Rad21-HA on *otr1* in the indicated strains treated (+) or not (-) with RNase. T, total extract; IP, immunoprecipitate. **d**, ChIP analyses of Ago1 on *otr1* relative to a transfer RNA gene in the indicated strains. WT, wild-type. **e**, Localization of HA-Ago1 and Cnp1 together at centromeres. DAPI, 4,6-diamidino-2-phenylindole.

Small RNA preparation and detection. Exponential-phase yeast cells were subjected to RNA extraction with TRIZOL (Invitrogen) and precipitation with propan-2-ol. Samples were resuspended in 50% formamide; 35 µg of total RNA was loaded on a 17.5% denaturing polyacrylamide gel containing 7 M urea. ³²P-labelled DNA probes complementary to centromeric *dh* repeats or the *ura4⁺ StuI/EcoRV* region were generated with a Random Primed DNA-labelling kit (Roche). The probes were hybridized to the membranes overnight at 42 °C in a rotating oven and washed twice with 2 × SSC, 2% SDS at 50 °C. Phosphor screens or films were exposed for a minimum of 3 h. DNA oligonucleotides of 24/20 nucleotides or 22 nucleotides in length complementary to the *dh* repeat or the *ura4⁺ StuI/EcoRV* region were used as markers.

Cytology. Immunofluorescence was performed as described³⁰, cells were fixed for 15–20 min in 3.7% freshly prepared formaldehyde for staining. The following antibodies were used: sheep anti-Cnp1 antiserum (dilution 1:300), mouse 12CA5 anti-HA (dilution 1:30). Secondary antibodies conjugated with fluorescein isothiocyanate (Sigma-Aldrich) or Alexa488 (Molecular Probes) were used at dilutions of 1:100 or 1:1,000. Microscopy was performed as described³⁰.

Received 9 February; accepted 22 April 2005.

Published online 19 June 2005.

- Hannon, G. J. & Rossi, J. J. Unlocking the potential of the human genome with RNA interference. *Nature* **431**, 371–378 (2004).
- Mello, C. C. & Conte, D. Jr. Revealing the world of RNA interference. *Nature* **431**, 338–342 (2004).
- Baulcombe, D. RNA silencing in plants. *Nature* **431**, 356–363 (2004).
- Jones, L., Ratcliff, F. & Baulcombe, D. C. RNA-directed transcriptional gene silencing in plants can be inherited independently of the RNA trigger and requires Met1 for maintenance. *Curr. Biol.* **11**, 747–757 (2001).
- Mette, M. F., Aufsatz, W., van der Winden, J., Matzke, M. A. & Matzke, A. J. Transcriptional silencing and promoter methylation triggered by double-stranded RNA. *EMBO J.* **19**, 5194–5201 (2000).
- Volpe, T. A. *et al.* Regulation of heterochromatic silencing and histone H3 lysine-9 methylation by RNAi. *Science* **297**, 1833–1837 (2002).
- Schramke, V. & Allshire, R. Hairpin RNAs and retrotransposon LTRs effect RNAi and chromatin-based gene silencing. *Science* **301**, 1069–1074 (2003).
- Lippman, Z. & Martienssen, R. The role of RNA interference in heterochromatic silencing. *Nature* **431**, 364–370 (2004).
- Mochizuki, K., Fine, N. A., Fujisawa, T. & Gorovsky, M. A. Analysis of a piwi-related gene implicates small RNAs in genome rearrangement in *Tetrahymena*. *Cell* **110**, 689–699 (2002).
- Mochizuki, K. & Gorovsky, M. A. Small RNAs in genome rearrangement in *Tetrahymena*. *Curr. Opin. Genet. Dev.* **14**, 181–187 (2004).
- Pal-Bhadra, M. *et al.* Heterochromatic silencing and HP1 localization in *Drosophila* are dependent on the RNAi machinery. *Science* **303**, 669–672 (2004).
- Kawasaki, H. & Taira, K. Induction of DNA methylation and gene silencing by short interfering RNAs in human cells. *Nature* **431**, 211–217 (2004).
- Morris, K. V., Chan, S. W., Jacobsen, S. E. & Looney, D. J. Small interfering RNA-induced transcriptional gene silencing in human cells. *Science* **305**, 1289–1292 (2004).
- Matzke, M. A. & Birchler, J. A. RNAi-mediated pathways in the nucleus. *Nature Rev. Genet.* **6**, 24–35 (2005).
- Martienssen, R. A. Maintenance of heterochromatin by RNA interference of tandem repeats. *Nature Genet.* **35**, 213–214 (2003).
- Grewal, S. I. & Moazed, D. Heterochromatin and epigenetic control of gene expression. *Science* **301**, 798–802 (2003).
- Orphanides, G. & Reinberg, D. A unified theory of gene expression. *Cell* **108**, 439–451 (2002).
- Proudfoot, N. New perspectives on connecting messenger RNA 3' end formation to transcription. *Curr. Opin. Cell Biol.* **16**, 272–278 (2004).
- Reed, R. Coupling transcription, splicing and mRNA export. *Curr. Opin. Cell Biol.* **15**, 326–331 (2003).
- Noma, K. *et al.* RITS acts in cis to promote RNA interference-mediated transcriptional and post-transcriptional silencing. *Nature Genet.* **36**, 1174–1180 (2004).
- Verdel, A. *et al.* RNAi-mediated targeting of heterochromatin by the RITS complex. *Science* **303**, 672–676 (2004).
- Partridge, J. F., Borgstrom, B. & Allshire, R. C. Distinct protein interaction domains and protein spreading in a complex centromere. *Genes Dev.* **14**, 783–791 (2000).
- Partridge, J. F., Scott, K. S., Bannister, A. J., Kouzarides, T. & Allshire, R. C. cis-acting DNA from fission yeast centromeres mediates histone H3 methylation and recruitment of silencing factors and cohesin to an ectopic site. *Curr. Biol.* **12**, 1652–1660 (2002).
- Motamedi, M. R. *et al.* Two RNAi complexes, RITS and RDRC, physically interact and localize to noncoding centromeric RNAs. *Cell* **119**, 789–802 (2004).
- Sugiyama, T., Cam, H., Verdel, A., Moazed, D. & Grewal, S. I. RNA-dependent RNA polymerase is an essential component of a self-enforcing loop coupling heterochromatin assembly to siRNA production. *Proc. Natl Acad. Sci. USA* **102**, 152–157 (2005).
- Aranda, A. & Proudfoot, N. Transcriptional termination factors for RNA polymerase II in yeast. *Mol. Cell* **7**, 1003–1011 (2001).
- Scafe, C., Nonet, M. & Young, R. A. RNA polymerase II mutants defective in transcription of a subset of genes. *Mol. Cell. Biol.* **10**, 1010–1016 (1990).
- Herr, A. J., Jensen, M. B., Dalmay, T. & Baulcombe, D. C. RNA polymerase IV directs silencing of endogenous DNA. *Science* **308**, 118–120 (2005).
- Onodera, Y. *et al.* Plant nuclear RNA polymerase IV mediates siRNA and DNA methylation-dependent heterochromatin formation. *Cell* **120**, 613–622 (2005).
- Pidoux, A. L., Richardson, W. & Allshire, R. C. Sim4: a novel fission yeast kinetochore protein required for centromeric silencing and chromosome segregation. *J. Cell Biol.* **161**, 295–307 (2003).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank members of the Allshire laboratory for advice, discussion and comments on the manuscript; E. Bayne for the model; P. Dunoyer and O. Voinnet for help with siRNA detection; T. Jenuwein for anti-H3K9me2 serum; and other members of the EC FP6 'The Epigenome' Network for input. This research was supported by funding from the Wellcome Trust to R.C.A., who is a Wellcome Trust Principal Research Fellow.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to R.C.A. (robin.allshire@ed.ac.uk).

CORRIGENDUM

doi:10.1038/nature03919

Mg isotope evidence for contemporaneous formation of chondrules and refractory inclusions

Martin Bizzarro, Joel A. Baker & Henning Haack

Nature 431, 275–278 (2004)

There is an error in our data-reduction algorithm, which resulted in a systematic bias in the ²⁷Al/²⁴Mg ratios published in this Letter by a factor of 1.11. Consequently, the (²⁶Al/²⁷Al)₀ values in the original Table 1 were underestimated (see corrected part of Table 1 below). The only significant change to our conclusions is that the initial ²⁶Al/²⁷Al at the time of formation of calcium–aluminium-rich inclusions (CAIs), as calculated from our data set, should be (5.83 ± 0.11) × 10^{–5}, which is in agreement with that for CAIs from carbonaceous chondrites¹. The main conclusions regarding the contemporaneous formation of some Allende chondrules and CAIs, as well as the brevity of the CAI-forming event, remain unchanged.

1. Young, E. D. *et al.* Supra-canonical ²⁶Al/²⁷Al and the residence time of CAIs in the solar protoplanetary disk. *Science* 308, 223–227 (2005).

Table 1 | Al-Mg isotopic data for Allende CAIs and chondrules

	²⁷ Al/ ²⁴ Mg	(²⁶ Al/ ²⁷ Al) ₀ (× 10 ^{–5})	ΔT ₀ (Myr)
CAIs			
A-1	0.224	5.85 ± 1.07	0.00 ^{–0.18} _{+0.21}
A-3c	1.34	5.89 ± 0.21	–0.011 ^{–0.037} _{+0.039}
A-7	2.66	5.74 ± 0.24	0.017 ^{–0.043} _{+0.045}
A-8a	3.42	5.77 ± 0.20	0.011 ^{–0.036} _{+0.037}
A-8b	2.49	5.79 ± 0.16	0.007 ^{–0.029} _{+0.030}
A-8c	3.44	5.82 ± 0.17	0.003 ^{–0.030} _{+0.031}
A-8d	2.95	6.02 ± 0.22	–0.035 ^{–0.038} _{+0.040}
A-8e	3.40	5.92 ± 0.17	–0.017 ^{–0.030} _{+0.031}
A-11	1.84	5.60 ± 0.19	0.042 ^{–0.035} _{+0.036}
A-13	1.83	5.85 ± 0.31	–0.003 ^{–0.054} _{+0.056}
Chondrules			
A-C1	0.780	2.34 ± 0.63	0.96 ^{–0.25} _{+0.33}
A-C1	0.973	1.51 ± 0.57	1.43 ^{–0.34} _{+0.51}
A-C2	0.928	5.68 ± 0.64	0.03 ^{–0.11} _{+0.13}
A-C5	0.295	1.84 ± 0.80	1.21 ^{–0.38} _{+0.60}
A-C14	0.023	–	–
A-C17	0.222	3.97 ± 1.10	0.41 ^{–0.25} _{+0.33}
A-C20	0.382	4.49 ± 0.63	0.28 ^{–0.14} _{+0.16}
A-C23	0.767	2.16 ± 0.55	1.04 ^{–0.24} _{+0.31}
A-C24	0.311	2.69 ± 0.77	0.81 ^{–0.26} _{+0.35}
A-C29	0.074	–	–
A-C31	0.277	4.49 ± 0.86	0.28 ^{–0.19} _{+0.22}
A-C32	0.097	–	–
A-C33	0.505	3.34 ± 0.47	0.59 ^{–0.14} _{+0.16}
A-B3	0.105	–	–
A-B6	0.247	6.05 ± 0.97	–0.04 ^{–0.16} _{+0.18}
A-B9	0.194	6.41 ± 1.23	–0.10 ^{–0.21} _{+0.27}
A-B9	0.217	6.23 ± 1.10	–0.07 ^{–0.17} _{+0.20}
A-B9	0.136	5.95 ± 1.70	–0.02 ^{–0.27} _{+0.37}

ERRATUM

doi:10.1038/nature03878

Large Cretaceous spenodontian from Patagonia provides insight into lepidosaur evolution in Gondwana

Sebastián Apesteguía & Fernando E. Novas

Nature 425, 609–612 (2003)

The Supplementary Information for this paper was not uploaded at the time of publication. It is now live at <http://www.nature.com/nature/journal/v425/n6958/full/nature01995.html>.

-  FOCUS
-  SPOTLIGHT
-  RECRUITMENT
-  ANNOUNCEMENTS
-  EVENTS

naturejobs

Take your partner by the hand...

Drug development is like a dance: it's difficult to do alone, and can be less satisfying. A report released last week shows that more biotechs are seeking partners for a bit of a bop. But pairing up with a partner entails some risks. When one half of a partnership dances a tango and the other a waltz, someone's toes are going to get trampled.

Critical Factors for Alliance Formation, from the consultancy company Deloitte, shows that the number of alliances forged between biotech companies more than doubled between 1999 and 2003. The biggest reason, of course, is that partnerships increase the availability of cash for research and development. But two key reasons relating to personnel rank close behind. Companies often need scientific expertise that they don't have in-house and they also need an effective sales and marketing team once they have a product.

Are all these new partnerships leading to redundancies, clearing the dance floor of jobs? Not necessarily. Biotech employment in the United States and Asia has been on the rise, although the sector has been a bit flat in Europe (see *Nature* **435**, 997; 2005). But it does mean that when

seeking a suitable biotech to work for, one should study not just the potential employer, but all its partners, then see which of your skills will help out everyone.

It also means that one skill in particular — intellectual property (IP) management — will be more in demand.

There was a time when biotechs mostly licensed out their technology and drug targets to pharmaceutical companies. Now they are dancing an IP two-step: sending some intellectual property out as well as bringing some in. This means there are potentially more jobs around for intellectual property managers — but at biotechs rather than at pharmaceutical companies.

Practising until you have a well-rounded skill set, both in science and in intellectual property, will allow you to face the music and dance — whatever size your partner.



Paul Smaglik, *Naturejobs* editor

CONTACTS

Publisher: Ben Crowe
Editor: Paul Smaglik
Marketing Manager: David Bowen

European Head Office, London
 The Macmillan Building, 4 Crinan Street
 London N1 9XW, UK
 Tel: +44 (0) 20 7843 4961
 Fax: +44 (0) 20 7843 4996
 e-mail: naturejobs@nature.com

Naturejobs Sales Director:
 Nevin Bayoumi (4978)
European Sales Manager:
 Andy Douglas (4975)

Natureevents Sales Manager:

Sille Opstrup (4994)
UK/RoW/Ireland/Italy:
 Nils Moeller (4953)
 Irene Viglia-Atton (4944)
Scandinavia/Spain/Portugal:
 Evelina Rubio Håkansson (4973)
France/Switzerland/Belgium:
 Amelie Pequignot (4974)
Germany/Austria/The Netherlands:
 Reya Silao (4970)

Advertising Production Manager:

Billie Franklin
 To send materials use London
 address above.

Tel: +44 (0) 20 7843 4814
 Fax: +44 (0) 20 7843 4996
 e-mail: naturejobs@nature.com
Naturejobs web development: Tom Hancock
Naturejobs online production: Niamh Shields

European Satellite Office
Germany/Austria/
The Netherlands:
 Patrick Phelan
 Tel: +49 89 54 90 57 11
 Fax: +49 89 54 90 57 20
 e-mail: p.phelan@nature.com

US Head Office, New York
 345 Park Avenue South,

10th Floor, New York,
 NY 10010-1707
 Tel: +1 800 989 7718
 Fax: +1 800 989 7103
 e-mail: naturejobs@natureny.com
US Sales Manager: Peter Bless

Japan Head Office, Tokyo
 Chiyoda Building,
 2-37 Ichigayatamachi,
 Shinjuku-ku,
 Tokyo 162-0843
 Tel: +81 3 3267 8751
 Fax: +81 3 3267 8746
Asia-Pacific Sales Director: Rinoko Asami
 e-mail: rasami@naturejpn.com

Immunology goes global

Scientists seeking immunology posts are looking beyond the United States and scattering all over the globe. They are re-evaluating both the focus of their work and where they choose to pursue it, says **Myrna Watanabe**.

Most of Rachel Kohler's Australian-born colleagues took posts in the United States at the first opportunity. But despite being offered immunology postdocs in New York and Boston, Kohler instead chose to stay in Australia.

The United States has a reputation for attracting the most prominent researchers, building the biggest facilities and providing the most ample funding. But things may be changing. As extramural funds from the National Institutes of Health (NIH) for supporting immunology research that isn't related to bioterror defence become scarcer, immunologists are shifting their job search away from the States. Other countries are fishing for immunologists who once considered the United States the top place to train, by offering sound packages and beefing up their infrastructure.

This sea change makes Kohler's decision to stay in Australia less surprising. "Within Australia, it is strongly impressed upon PhD students that an overseas postdoc is paramount to one's career," says Kohler, who completed her degrees at the University of Adelaide and works at the Centenary Institute of Cancer Medicine and Cell Biology in Sydney. Her friends in the field, she notes, are all working overseas.

Kohler's situation is no aberration, says Robert Kimberly, professor of medicine and director of clinical immunology and rheumatology at the University of Alabama, Birmingham. He has seen a tidal shift from ten years ago, when many US immunology postdocs came from Western Europe. But now, as the immunology infrastructure has strengthened in Europe, international postdocs in the United States are coming from elsewhere.

Wherever they end up, immunology postdocs have a broad field from which to choose. Immunology research ranges from animal models of infectious diseases to human microRNAs and their effects on gene expression. For this reason, immunologists like Kimberly who have spent decades in the discipline, now recommend a less specialized approach to training.

Iqbal Grewal, vice-president for preclinical therapeutics at Seattle Genetics, a biotech firm in Seattle, Washington, suggests a broad background: both basic and applied immunology, plus knowledge of the human disease process, the human genome, molecular biology, and perhaps a PhD in immunology or an MD/PhD.

Neetu Gupta has just such a multidisciplinary and international background. Gupta, who is spending her second postdoc at the University of California, San Francisco, did an undergraduate degree in

IMAGE
UNAVAILABLE
FOR COPYRIGHT
REASONS

Meet the competition:
immunology labs around
the world are increasingly
competing on equal terms
with their US counterparts.

zoology, a master's in biotechnology and a PhD in immunochemistry. She spent her first postdoc working for the NIH at the immunogenetics lab of the National Institute of Allergy and Infectious Diseases (NIAID).

This approach helped set the next stage of her career. She has a mentored career development award from the NIH, which she can use to secure her next position. She plans to look for a job in the autumn. "I have a grant and this grant is portable," Gupta says.

Tarek Fahmy, assistant professor of biomedical engineering at Yale, has one foot in industry and the other in academia, and illustrates how multidisciplinary training can provide a robust career in immunology. Fahmy first trained as a chemical engineer and then worked for DuPont for four years. While with the company, he reassessed his career goals and enrolled in an immunology graduate programme at Johns Hopkins University in Baltimore, Maryland. His decision to enter immunology was personal. "My father had died of lymphoma," he says. He is now founding a biotech that aims to use immunology to fight cancer.

Perhaps because fewer European scientists are launching their immunology careers in the United States, more positions seem to be available, with relatively low competition for them.

William Paul, chief of the immunology laboratory at the NIAID, says the NIH employs 500 or more people in immunology, about a third of them in the NIAID. While recruiting for two tenure-track investigator positions (the equivalent of assistant professors), Paul was surprised that, despite the positions being funded, he received a mere 56 applications. He eventually sent 350 more e-mails soliciting candidates. His short-list of

seven now contains three who were either born in the United States or spent most of their lives there.

There could be several reasons for this trend. First, although the NIH's budget doubled a few years ago, its funding plateau since then means there is less money for new grants. And immunology faces even more of a crunch. Of the new money that is available, most has been earmarked for research relating to biowarfare and defence, leaving other research out in the cold. Second, researchers from Asia are having a harder time getting US visas, and are wary of coming as they have heard horror stories about colleagues who were refused re-entry or had difficulty travelling to and from the states. In immunology, getting a visa may be even harder for those working in areas remotely related to biodefence, because of security concerns, real or imagined.

Mark Buller, professor of molecular microbiology and immunology at St Louis University in Missouri, has a new postdoc coming from the University of Cambridge in Britain later this year. Now he is looking for another postdoc with a virology background to work on the mechanism underlying a poxvirus model. Most of his postdoc applications come from India and China, but because he has had difficulty securing visas for prospective fellows from these countries in the past few years, he is less likely to recruit them now.

Recruiting staff is even harder for so-called biosafety level 3 and 4 (BSL-3 and BSL-4) labs — highly secure labs that often handle dangerous pathogens such as smallpox. These labs have a different set of requirements from others, according to Lynn Soong of the Center for Biodefense and Emerging Infectious Diseases at the University of Texas Medical Branch in Galveston, the site of a new BSL-4 facility. Besides standard immunology training, candidates need to have experience with selective agents of disease, such as West Nile virus, Lassa fever and rickettsia. People who handle these agents in BSL-4 laboratories must also be prepared to work independently. Because of the

OTHER WORLD-CLASS CENTRES

Universities in Europe and around the world that once lost many a recruiting competition with their US brethren are now winning, through growth and infrastructure investment. Most tout their collaborative credentials, including joint programmes with universities inside and outside their own countries, and with researchers from other countries.

The Marseille-Luminy Immunology Centre, for example, not only collaborates with French research institutions, but also works with others elsewhere in Europe. And the Pasteur Institute in Paris has several departments involved in immunology, with the Pasteur Foundation supporting postdoctoral fellowships specifically for US citizens.

In Britain, the Division of Immunology, Infection and Inflammation at the University of Glasgow runs highly collaborative programmes investigating immune regulation and development, and virulence factors in bacteria.

Meanwhile, the Division of Infections

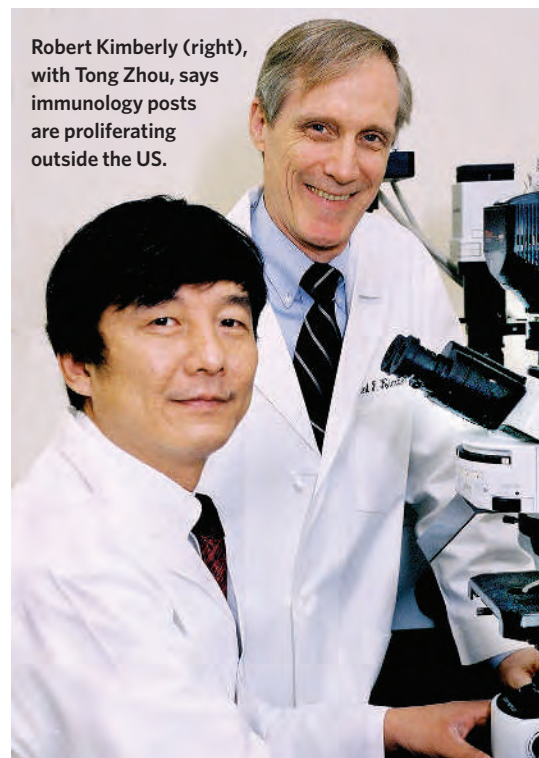
and Immunity at the National Institute for Medical Research in Mill Hill concentrates on immune development, the molecular mechanisms of intracellular infection, and the effects of vaccination.

The Institute for Research in Biomedicine in Bellinzona, Switzerland, was founded in 2000 and already has 10 research groups focusing on molecular and cellular immunology.

On the other side of the globe, in Australia, the Walter and Eliza Hall Medical Research Institute in Parkville, near Melbourne, has a long history of immunology research and currently focuses on the function and development of the immune system, including autoimmunity and transplantation.

In Japan, the Institute of Physical and Chemical Research (RIKEN) established a Research Center for Allergy and Immunology in Yokohama in 2001. The institute has 27 laboratories and research units, focusing on everything from cell signalling to allergies. **M.W.**

CROWN COPYRIGHT/HEALTH & SAFETY LAB/SPL



Robert Kimberly (right), with Tong Zhou, says immunology posts are proliferating outside the US.

WEB LINKS

Division of Immunology, Infection and Inflammation, University of Glasgow

♦ www.gla.ac.uk/departments/immunology
RIKEN, Japan

♦ www.riken.jp/engn
Institute of Infectious Disease and Molecular Medicine, University of Cape Town

♦ www.iidmm.uct.ac.za
Walter and Eliza Hall Medical Research Institute, Australia

♦ www.wehi.edu.au
Centenary Institute, Sydney, Australia

♦ www.mja.com.au/public/issues/171_11_061299/basten/basten.html
Marseille-Luminy Immunology Centre

♦ www.ciml.univ-mrs.fr

high security, they will log into the lab, work for several hours, and then log out — they can't just run out to grab a senior researcher to show them something.

Meanwhile, more job opportunities at new or growing institutes outside the United States are creating competition for global talent. Australia's Centenary Institute, besides actively recruiting postdocs, is looking for a new head. Nick Pearce, Centenary's business development manager, says the institute will specifically target Australian expatriates.

The Institute of Infectious Disease and Molecular Medicine at the University of Cape Town in South Africa is also looking for a director. According to Frank Brombacher, a professor of immunology who works on knock-out mouse models, the facility receives national and international funding, with some coming from the Wellcome Trust in Britain. Brombacher, who hails from Germany and received his PhD from the Max Planck Institute for Immunobiology in Freiburg am Breisgau, is looking for someone to fill a lecturer's position, which is 90% research and 10% teaching. He is also looking for up to two postdocs to study murine models of infectious diseases.

In Singapore, the Biopolis initiative (see *Nature* 425, 746–747; 2003) is producing many opportunities for immunologists, says Kong Peng Lam, acting executive director of the Biomedical Research Council of Singapore's Agency for Science, Technology and Research (A*STAR). The National University of Singapore is also actively recruiting immunologists. Lam, whose undergraduate and graduate training was in the United States and who did a postdoc in Germany, returned to Singapore in 1998 to join the Institute of Molecular and Cell Biology.

As more institutions like Biopolis rise and grow, so will the global competition for immunologists. Broad training — and a willingness to travel globally — will help young immunologists to find jobs in an exciting multidisciplinary field.

Myrna Watanabe is a freelance writer based in Patterson, New York.

Are we not men?

Meet the family...

Henry Gee

After that they started popping up all over the place. Not that this was always advisable. Sometimes they were shot by loggers even before they could catch colds from Discovery Channel camera crews.

Sometimes they ran into one another. A press conference given by a group of media-savvy pygmy indigenes from Northern Sulawesi was disrupted when a rival group of hitherto unknown hominids of enormous size ate the pygmies and ran off with the A/V equipment.

And sometimes they just tripped over their own feet. Like the centuries-old Alma chieftain who admitted (on live, prime-time TV, and in rounded Oxford tones) how much he liked Tolkien, and went on to describe in toothsome detail the sado-masochistic sexual cannibalism at the heart of Yeti religion. Post-modernist chatterati were left in agonies of indecision about which solecism was worse.

What was so remarkable was how soon the fuss died down. It was as if the hominids had been waiting for the right time to emerge from their fastnesses, a time when *Homo sapiens* wouldn't automatically seek to destroy them. That after our own sorrows — the abandonment of much of Africa in the 2020s due to AIDS and famine, and the haemorrhagic plagues that killed one in three people in the 2030s — we were now mature company for any self-respecting species on Earth.

When the time came, they just settled down with us, side by side. Just ten years after the first Sasquatches came out of northern British Columbia in '39 in search of whiskey, the hominids were everywhere, and nobody raised a brow-ridge. It would be commonplace to find (say) a Sumatran Pendek driving your cab to work; your lunch cooked and served by a Malaysian Jive Monkey (and before you complain, that's what they called *themselves*); and an eight-foot Kaptar from the Pamirs, pole-dancing to Earth, Wind and Fire in a club after work (but only if you were into that kind of thing).

But this acceptance came at a cost. Many of us continued to assume that we (or 'We', or 'People', or 'HomSap') were a breed apart. And so we were: just one among 20 or so species of hominid, and by far the most numerous. But what some of the remnants of religion could not stomach was that we were no longer The Elect, The Chosen. These remnants were small, but vocal. But who were they? The Muslims had long since decreed that the woes of mankind were the will of Allah, and that was that. The Catholics were, well, catholic, and in a famous encyclical, *Undique humanitas*, Pope Eusebius decreed that all hominids were ensouled creatures

Kingdom of Granada in 1492, their pretext (it turned out) was that the Emir had had 'Devils' as bodyguards. When we finally got to the sub-cellars beneath the Alhambra, we found them — the great, hulking bones of classic Neanderthals. And we could take their DNA.

Once we thought that Neanderthals weren't closely related to Us. But the Neanderthals used in ancient-DNA studies were Ice Age examples from long ago. We had never seen DNA from Neanderthals living so recently. And all of a sudden it made sense — the reason why Clovis was Hairy; the big noses and brooding, beetling expressions of everyone from — say — Leonardo to Einstein.



JACEY

of God. The Jews welcomed the opportunity for God to choose someone else for a change. The last hold-outs were sabbatarian enclaves in the United States and parts of West Africa who refused to countenance that the hominids were really human — the first for reasons of racial superiority, the second so as not to disturb the bush-meat trade.

It took one event to convince everybody. No, it wasn't when Serumthrep Okk, an Alma from the Altai, was declared the next Incarnation of the Holy One. And not even when one Jjkkaa'HhkhHoj, millionaire scion of a Jacksonville rental car business, became the first Tibestian Sand-Druid to be Bar-Mitzvah (*mazel tov*). But it came from an echo of the past.

There's nothing new under the sun, you see, for we'd met hominids before. Those fairy stories were firmly based in fact. When Ferdinand and Isabella invaded the

After that they started popping up all over the place. Abraham Lincoln had been at least 35% Sasquatch. Most of the Khmer Rouge had been Malaysian Jive Monkeys. (I still can't believe that name. But they're a fun crowd.) The final knell came when it was announced that Charles Darwin had been more than 65% Neanderthal, a value that turned out to be typical of British aristocracy, exceeded only by the immediate parentage of (you guessed it) His Holiness, Pope Eusebius, whose family had lived in southern Spain since time immemorial.

With typical political aplomb, the Pope had been ahead of the game all the time. Now we'll all have to get used to it. There are hominids in us all. ■

Henry Gee is a senior editor of *Nature*. His latest book is *The Science of Middle Earth* (Cold Spring/Souvenir).